YOVA KEMENTCHEDJHIEVA

METHODS, EVALUATIONS AND RESOURCES FOR MULTILINGUAL TRANSFER LEARNING

METHODS, EVALUATIONS AND RESOURCES FOR MULTILINGUAL TRANSFER LEARNING

yova kementchedjhieva

PhD Thesis

February 2021

THESIS SUPERVISOR:
Anders Søgaard

ASSESSMENT COMMITTEE:
Christian Igel, University of Copenhagen
Mirella Lapata, University of Edinburgh
Mikel Artetxe, DeepMind

AFFILIATION:
Department of Computer Science
Faculty of Science
University of Copenhagen

Yova Kementchedjhieva: *Methods, Evaluations and Resources for Multilingual Transfer Learning,* February 2021

ABSTRACT

Language technology has transformed the way we write, the way we interact with our devices, and the way we share and consume information. This was made possible by advancements in the field of Natural Language Processing (NLP), a largely data-driven subfield of machine learning. Since data are limited for many of the tasks, domains and languages studied in NLP, transfer learning has gained great prominence in the field as a way to alleviate data scarcity. This thesis presents work on methods, evaluations and resources for multilingual transfer learning. Our research shows how to improve and correctly evaluate cross-lingual embeddings obtained through alignment. It sheds light on the source of performance in cross-lingual transfer learning for dependency parsing. And it introduces two new resources for language generation tasks, one best viewed as a test bed for cross-domain transfer methods and the other, as a test bed for meta-learning techniques. This thesis contributes to efforts in NLP towards optimal transfer of knowledge across languages and highlights some remaining limitations.

ABSTRACT IN DANISH – ABSTRAKT PÅ DANSK

Sprogteknologi har transformeret den måde vi skriver på, den måde vi interagerer med vores digitale enheder på og den måde vi deler og forbruger information på. Denne transformation er muliggjort som følge af fremskridt inden for Natural Language Processing (NLP), et hovedsageligt datadrevet underfelt indenfor maskinlæring. Da data er en begrænset ressource der anvendes til mange af de opgaver, domæner og sprog der studeres i NLP, har overførselslæring fået fremtrædende plads i feltet som en måde at lindre dataknaphed på. Denne afhandling præsenterer metoder, evalueringer og ressourcer der kan anvendes til flersproget *tranfer learning*. Den udarbejdede forskning viser, hvordan man kan forbedre og korrekt evaluere flersprogede indlejringer opnået gennem matchning. Ydermere kastes der lys over grundlaget for ydeevnen i flersproget *tranfer learning* for *dependency parsing*. Derudover introducerer afhandlingen to nye ressourcer til sproggenereringsopgaver, den ene ses bedst som en testplatform til transfer metoder på tværs af domæner og den anden som en testplatform til metalæringsteknikker. Denne afhandling bidrager til indsatsen i NLP mod optimal overførsel af viden på tværs af sprog og fremhæver nogle resterende begrænsninger.

PUBLICATIONS

This is an article-based thesis. Chapters 2 to 7 each represent a peer-reviewed article. The articles are identical in content as they appear here and in the original publications, except for minor changes such as the correction of typos and reformatting of tables and figures. The following articles are included in the thesis:

Hartmann, Mareike, Yova Kementchedjhieva, and Anders Søgaard (2019). „Comparing Unsupervised Word Translation Methods Step by Step." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32.

Kementchedjhieva, Yova, Mareike Hartmann, and Anders Søgaard (Nov. 2019). „Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary Induction." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 3336–3341.

Kementchedjhieva, Yova, Di Lu, and Joel Tetreault (Dec. 2020). „The ApposCorpus: a new multilingual, multi-domain dataset for factual appositive generation." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online), pp. 1989–2003.

Kementchedjhieva, Yova, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard (Oct. 2018). „Generalizing Procrustes Analysis for Better Bilingual Dictionary Induction." In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium, pp. 211–220.

Şahin, Gözde Gül, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych (July 2020). „PuzzLing Machines: A Challenge on Learning From Small Data." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, pp. 1241–1254.

Vania, Clara, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez (Nov. 2019). „A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 1105–1116.

I was also involved in the following publications that are not included in this thesis:

Barrett, Maria, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard (Nov. 2019). „Adversarial Removal of Demographic Attributes Revisited." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 6330–6335.

Bjerva, Johannes, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein (June 2019a). „A Probabilistic Generative Model of Linguistic Typology." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota, pp. 1529–1540.

Bjerva, Johannes, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein (July 2019b). „Uncovering Probabilistic Implications in Typological Knowledge Bases." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 3924–3930.

Gonen, Hila, Yova Kementchedjhieva, and Yoav Goldberg (Nov. 2019). „How Does Grammatical Gender Affect Noun Representations in Gender-Marking Languages?" In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China, pp. 463–471.

Hartmann, Mareike, Yova Kementchedjhieva, and Anders Søgaard (2018). „Why is unsupervised alignment of English embeddings from different algorithms so hard?" In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 582–586.

Kementchedjhieva, Yova, Johannes Bjerva, and Isabelle Augenstein (Oct. 2018). „Copenhagen at CoNLL–SIGMORPHON 2018: Multilingual Inflection in Context with Explicit Morphosyntactic Decoding." In: *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*. Brussels, pp. 93–98.

Kementchedjhieva, Yova and Adam Lopez (Nov. 2018). „'Indicatements' that character language models learn English morpho-syntactic units and regularities." In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium, pp. 145–153.

Ruder, Sebastian, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard (2018). „A Discriminative Latent-Variable Model for Bilingual Lexicon Induction." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 458–468.

ACKNOWLEDGEMENTS

I am sad to see this journey come to an end because against all odds doing a PhD has been a wonderful experience for me. For that I have to thank first and foremost my research group CoAStaL, which is (and I say this with 100% certainty) the *best* collection of human beings: warm, kind, open-minded, fun-loving, and caring—and they also happen to like NLP, so along all the good memories we made, we also published some papers. The role of my supervisor Anders Søgaard in putting this group together is commendable. Thank you, Anders, for giving me the opportunity to be a part of CoAStaL, for teaching me through words and example about the importance of work-life balance, about the value of mindfulness, and of course about NLP research.

I thank all my collaborators for giving me the opportunity to learn from them—I hope we will continue to work together. Special thanks to Adam Lopez, formerly my Master's thesis supervisor, for his continued guidance, always delivered with so much thought and patience.

Some friends who deserve buckets of gratitude for all the support they have given me throughout the PhD include Mareike: colleague, friend, flatmate, and somehow even more than the sum of these things; Desmond: a role model and a dear friend; Vickie: my high-school partner in crime and a friend for life; and Mark: my favourite person in the whole wide world.

And of course a big thank you to my mother for making it possible for me to be where I am today, for calling me out on my fear of maths, and for her constant support and encouragement. I hope that this achievement compensates at least in part for being so far from home.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Language technology has become integral to the digital activities one carries out on a daily basis. That is, if one carries out their daily activities in English, or in one of a few other languages of similarly high socioeconomic status. For the purposes of Natural Language Processing (NLP), the machine learning subfield behind all language technology we know, this status translates into high data availability. That in turn enables the development of a range of high-quality functionalities for these *high-resource* languages, like getting assistance in writing emails, easily finding relevant content online, making appointments with voice control, discovering the best restaurants, movies, books for one's taste, and many more. Meanwhile, the development of language technology for the other several thousand languages spoken in the world is lagging behind for two main reasons: a data scarcity ranging in magnitude from merely inconvenient to largely insurmountable, and a longstanding lack of an incentive for researchers in both industry and academia to work outside of the few 'chosen' languages. The last few years have seen a shift in attitudes with respect to the latter, motivated by goals of equality and democratization of the Internet and its associated technology. The matter remains, however, that data for many of the world's languages are limited, rendering most of the methods designed for high-resource languages unusable in the general case.

Consider as an example the spell-checking function of Google Docs, the widely used online word processor, and how it responds to the same sentence when presented in English, Bulgarian and Macedonian. In all cases the sentence contains two errors, a grammatical one and a typographical one:

He speak to the manageer.
Той говоря с мениджъьра.
Тој зборувам со менаџеерот.

In Google Docs a blue wavy line signifies bad grammar and a red wavy line signifies a typo. In the English sentence, we see the incorrect verb form marked as a grammatical error and the typo in *manager* marked as such. In Bulgarian, on the other hand, the word processor can detect typos but not grammatical errors. Bulgarian is a language of only about nine million speakers, but it is also one of the 24 official languages of the European Union.[1] Meanwhile, Macedonian, which has not yet obtained this status and is spoken by only about two

[1] The proceedings of the European Parliament are a great source of language data.

million people, is simply not supported by Google's spelling and grammar tools.

The future of language technology for Bulgarian, Macedonian and the numerous even less resourceful languages of the world lies with *multilingual* NLP, a branch of NLP which focuses on the development of algorithms that work for languages with very different linguistic properties and that, crucially, can work in settings of limited data availability. The latter requirement motivates the development of a range of techniques that fall under the scope of *transfer learning*. They aim to alleviate the problems that arise in low-resource settings by leveraging data, i.e. transferring knowledge, from other domains, tasks, and languages.

TRANSFER LEARNING is a broad term that refers to training NLP models on data other than those available for the specific task one wants to solve (Pan and Yang, 2010). One example of a *target task, domain* and *language* is grammatical error correction (GEC) for formal documents in Macedonian. If insufficient data are available for this specific task, one could resort to transfer learning (a) across domains, if GEC data are available for other types of text in Macedonian, e.g. for Wikipedia articles, (b) across tasks, if data are available for another task that is related to grammar e.g. part-of-speech tagging, and (c) across languages, if GEC data are available for another language. In many cases data would be available in multiple other languages, and the choice among them can be very consequential. Cross-lingual transfer takes place more naturally between languages with similar properties, which is often the case for languages from the same family. For Macedonian, for example, a good *source* language for transfer learning could be Bulgarian, a closely related Slavic language with similar vocabulary and grammar.

Transfer of knowledge, whether it is across domains, tasks or languages, can happen in three main ways: via the pre-training of an NLP model on source data, optionally followed by fine-tuning on target data or used directly to solve the target task in a zero-shot fashion; via multi-task, cross-domain or cross-lingual training, which all refer to the training of a model simultaneously on source and target data across tasks, domains or languages, respectively; or via meta-learning of the initial parameters of an NLP model—in this case the knowledge being transferred does not pertain to the specific task at hand, but rather to the general skill of learning to solve a task from limited evidence.

THIS THESIS presents work on the development of methods and resources for transfer learning in multilingual, low-resource settings.

The first three chapters concern the topic of cross-lingual word embedding (CLWE) alignment. Chapter 2 presents a supervised method for the alignment of English to a low-resource language improved via anchoring to a third high-resource language related to the low-resource language of interest. Chapter 3 argues for a fair comparison of unsupervised methods for cross-lingual embedding alignment, showing that their separate components need to be set side by side in a controlled setting. And Chapter 4 shows that one key benchmark used in the evaluation of cross-lingual embeddings on the task of bilingual dictionary induction is deeply flawed.

Chapter 5 measures the individual and combined merit of data augmentation and cross-lingual transfer learning for dependency parsing in *extremely* low-resource settings.

Chapter 6 presents a multi-lingual, multi-domain dataset for the little studied task of appositive generation, a task which we argue is best approached through transfer learning, due to the rare occurrence of appositives in text which results in data sparsity.

In Chapter 7 we introduce a challenge dataset based on language puzzles from Linguistic Olympiads. This dataset tests the abilities of NLP models to learn to translate from one language to another based on as few as ten parallel sentences.

2 GENERALIZING PROCRUSTES ANALYSIS FOR BETTER BILINGUAL DICTIONARY INDUCTION

ABSTRACT

Most recent approaches to bilingual dictionary induction find a linear alignment between the word vector spaces of two languages. We show that projecting the two languages onto a third, latent space, rather than directly onto each other, while equivalent in terms of expressivity, makes it easier to learn approximate alignments. Our modified approach also allows for supporting languages to be included in the alignment process, to obtain an even better performance in low resource settings.

2.1 INTRODUCTION

Several papers recently demonstrated the potential of very weakly supervised or entirely unsupervised approaches to bilingual dictionary induction (BDI) (Artetxe, Labaka, and Agirre, 2017; Barone, 2016; Conneau et al., 2018; Søgaard, Ruder, and Vulić, 2018; Zhang et al., 2017), the task of identifying translational equivalents across two languages. These approaches cast BDI as a problem of aligning monolingual word embeddings. Pairs of monolingual word vector spaces can be aligned without any explicit cross-lingual supervision, solely based on their distributional properties (for an adversarial approach, see Conneau et al. (2018)). Alternatively, weak supervision can be provided in the form of numerals (Artetxe, Labaka, and Agirre, 2017) or identically spelled words (Søgaard, Ruder, and Vulić, 2018). Successful unsupervised or weakly supervised alignment of word vector spaces would remove much of the data bottleneck for machine translation and push horizons for cross-lingual learning (Ruder, Vulić, and Søgaard, 2018).

In addition to an unsupervised approach to aligning monolingual word embedding spaces with adversarial training, Conneau et al. (2018) present a supervised alignment algorithm that assumes a gold-standard seed dictionary and performs Procrustes Analysis (Schönemann, 1966). Søgaard, Ruder, and Vulić (2018) show that this approach, weakly supervised with a dictionary seed of *cross-lingual homographs*, i.e. words with identical spelling across source and target language, is superior to the completely unsupervised approach. We therefore focus on weakly-supervised Procrustes Analysis (PA) for BDI here.

The implementation of PA in Conneau et al. (2018) yields notable improvements over earlier work on BDI, even though it learns a simple linear transform of the source language space into the target language space. Seminal work in supervised alignment of word vector spaces indeed reported superior performance with linear models as compared to non-linear neural approaches (Mikolov, Le, and Sutskever, 2013). The relative success of the simple linear approach can be explained in terms of isomorphism across monolingual semantic spaces,[1] an idea that receives support from cognitive science (Youn et al., 2016). Word vector spaces are not *perfectly* isomorphic, however, as shown by Søgaard, Ruder, and Vulić (2018), who use a Laplacian graph similarity metric to measure this property. In this work, we show that projecting both source and target vector spaces into a *third* space (Faruqui and Dyer, 2014), using a variant of PA known as Generalized Procrustes Analysis (Gower, 1975), makes it easier to learn the alignment between two word vector spaces, as compared to the single linear transform used in Conneau et al. (2018).

CONTRIBUTIONS We show that Generalized Procrustes Analysis (GPA) (Gower, 1975), a method that maps two vector spaces into a third, latent space, is superior to PA for BDI, e.g., improving the state-of-the-art on the widely used English-Italian dataset (Dinu, Lazaridou, and Baroni, 2015) from a P@1 score of 66.2% to 67.6%. We compare GPA to PA on aligning English with five languages representing different language families (Arabic, German, Spanish, Finnish, and Russian), showing that GPA consistently outperforms PA. GPA also allows for the use of additional support languages, aligning three or more languages at a time, which can boost performance even further. We present experiments with multi-source GPA on an additional five low-resource languages from the same language families (Hebrew, Afrikaans, Occitan, Estonian, and Bosnian), using their bigger counterpart as a support language. Our code is publicly available.[2]

2.2 PROCRUSTES ANALYSIS

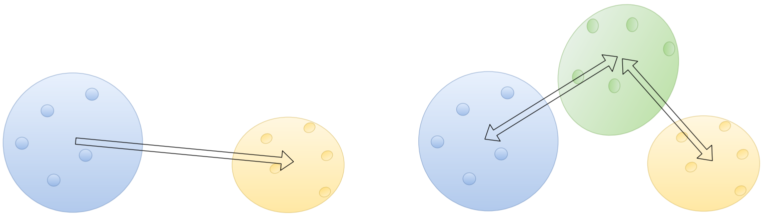
Procrustes Analysis is a graph matching algorithm, used in most mapping-based approaches to BDI (Ruder, Vulić, and Søgaard, 2018). Given two graphs, E and F, Procrustes finds the linear transformation T that minimizes the following objective:

$$\arg\min_{T} ||TE - F||^2 \tag{2.1}$$

thus minimizing the trace between each two corresponding rows of the transformed space TE and F. We build E and F based on a seed

[1] Two vector spaces are isomorphic if there is an invertible linear transformation from one to the other.

[2] `https://github.com/YovaKem/generalized-procrustes-MUSE`



(a) Procrustes Analysis (b) Generalized Procrustes Analysis

Figure 2.1: Visualization of the difference between PA, which maps the source space directly onto the target space, and GPA, which aligns both source and target spaces with a third, latent space, constructed by averaging over the two language spaces.

dictionary of N entries, such that each pair of corresponding rows in E and F, (e_n, f_n) for $n = 1, \ldots, N$ consists of the embeddings of a translational pair of words. In order to preserve the monolingual quality of the transformed embeddings, it is beneficial to use an orthogonal matrix T for cross-lingual mapping purposes (Artetxe, Labaka, and Agirre, 2017; Xing et al., 2015).[3] Conveniently, the orthogonal Procrustes problem has an analytical solution, based on Singular Value Decomposition (SVD):

$$\begin{aligned} F^\top E &= U\Sigma V^\top \\ T &= VU^\top \end{aligned} \tag{2.2}$$

2.3 GENERALIZED PROCRUSTES ANALYSIS

Generalized Procrustes Analysis (Gower, 1975) is a natural extension of PA that aligns k vector spaces at a time. Given embedding spaces $E_1, \ldots, E_k$, GPA minimizes the following objective:

$$\arg \min_{\{T_1, \ldots, T_k\}} \sum_{i<j}^{k} ||T_i E_i - T_j E_j||^2 \tag{2.3}$$

For an analytical solution to GPA, we compute the average of the embedding matrices $E_{1\ldots k}$ after transformation by $T_{1\ldots k}$:

$$G = k^{-1} \sum_{i=1}^{k} E_i T_i \tag{2.4}$$

[3] Recently, Doval et al. (2018) showed that the monolingual quality of embeddings need not suffer from a transformation guided by cross-lingual alignment, but their method still relies on an initial alignment obtained e.g. with Procrustes analysis, as described here.

thus obtaining a latent space, G, which captures properties of each of $E_{1\ldots k}$, and potentially additional properties emerging from the combination of the spaces. On the very first iteration, prior to having any estimates of $T_{1\ldots k}$, we set $G = E_i$ for a random i. The new values of $T_{1\ldots k}$ are then obtained as:

$$\begin{aligned} G^\top E_i &= U\Sigma V^\top \\ T_i &= VU^\top \text{ for } i \text{ in } 1\ldots k \end{aligned} \tag{2.5}$$

Since G is dependent on $T_{1\ldots k}$ (see Eq.2.4), the solution of GPA cannot be obtained in a single step (as is the case with PA), but rather requires that we loop over subsequent updates of G (Eq.2.4) and $T_{1\ldots k}$ (Eq.2.5) for a fixed number of steps or until satisfactory convergence. We observed little improvement when performing more than 100 updates, so we fixed that as the number of updates.

Notice that for $k = 2$ and with the orthogonality constraint in place, the objective for Generalized Procrustes Analysis (Eq. 2.3) reduces to that for simple Procrustes (Eq. 2.1):

$$\begin{aligned} &\arg\min_{\{T_1,T_2\}} ||T_1E_1 - T_2E_2||^2 \\ &= \arg\min_T ||TE_1 - E_2||^2 \\ &\qquad \text{where } T = T_1T_2^T \end{aligned} \tag{2.6}$$

Here T itself is also orthogonal. Yet, the solution found with GPA may differ from the one found with simple Procrustes: the former maps E_1 and E_2 onto a third space, G, which is the average of the two spaces, instead of mapping E_1 directly onto E_2. To understand the consequences of this difference, consider a single step of the GPA algorithm where after updating G according to Eq.2.4 we are recomputing T_1 using SVD. Due to the fact that G is partly based on E_1, these two spaces are bound to be more similar to each other than E_1 and E_2 are.[4] Finding a good mapping between E_1 and G, i.e. a good setting of T_1, should therefore be easier than finding a good mapping from E_1 to E_2 directly. In this sense, by mapping E_1 onto G, rather than onto E_2 (as PA would do), we are solving an easier problem and reducing the chance of a poor solution.

2.4 EXPERIMENTS

In our experiments, we generally use the same hyper-parameters as used in Conneau et al. (2018), unless otherwise stated. When extracting dictionaries for the bootstrapping procedure, we use cross-domain local scaling (CSLS, see Conneau et al. (2018) for details) as a metric for ranking candidate translation pairs, and we only use the ones that rank higher than 15,000. We do not put any restrictions

[4] A theoretical exception being the case there E_1 and E_2 are identical.

High-resource	AR	DE	ES	FI	RU
	575k	2,183k	1,412k	437k	1,474k
Low-resource	HE	AF	OC	ET	BS
	224k	49k	84k	175k	77k

Table 2.1: Statistics for Wikipedia corpora.

	AR		DE		ES		FI		RU		ave	
	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$
PA	34.73	61.87	73.67	91.73	81.67	92.93	45.33	75.53	47.00	79.00	56.48	80.21
GPA	**35.33**	**64.27**	**74.40**	**91.93**	**81.93**	**93.53**	**47.87**	**76.87**	**48.27**	**79.13**	**57.56**	**81.15**

Table 2.2: Bilingual dictionary induction performance, measured in P@k, of PA and GPA across five language pairs.

on the initial seed dictionaries, based on cross-lingual homographs: those vary considerably in size, from 17,012 for Hebrew to 85,912 for Spanish. Instead of doing a single training epoch, however, we run PA and GPA with early stopping, until five epochs of no improvement in the validation criterion as used in Conneau et al. (2018), i.e. the average cosine similarity between the top 10,000 most frequent words in the source language and their candidate translations as induced with CSLS. Our metric is Precision at k×100 (P@k), i.e. percentage of correct translations retrieved among the *k* nearest neighbor of the source words in the test set (Conneau et al., 2018). Unless stated otherwise, experiments were carried out using the publicly available pre-trained fastText embeddings, trained on Wikipedia data,[5] and bilingual dictionaries—consisting of 5000 and 1500 unique word pairs for training and testing, respectively—provided by Conneau et al. (2018)[6].

2.4.1 *Comparison of PA and GPA*

HIGH RESOURCE SETTING We first present a direct comparison of PA and GPA on BDI from English to five fairly high-resource languages: Arabic, Finnish, German, Russian, and Spanish. The Wikipedia corpus sizes for these languages are reported in Table 2.1. Results are listed in Table 2.2. GPA improves over PA consistently for all five languages. Most notably, for Finnish it scores 2.5% higher than PA.

COMMON BENCHMARKS For a more extensive comparison with previous work, we include results on English–{Finnish, German, Italian} dictionaries used in Conneau et al. (2018) and in Artetxe,

[5] https://github.com/facebookresearch/fastText
[6] https://github.com/facebookresearch/MUSE

	IT			DE			FI		
	5000	Identical	Num.	5000	Identical	Num.	5000	Identical	Num.
				Wacky					
VecMap	45.27*	38.33	39.40*	44.27*	40.73	40.27*	32.94*	27.39	26.47*
PA	44.90	45.47	01.13	47.26	47.20	45.93	**33.50**	**31.46**	01.05
GPA	**45.33**	**45.80**	**45.93**	**48.46**	**47.60**	**47.60**	31.39	31.04	**28.93**
				Wikipedia					
PA	66.24	66.39	-	65.33	64.77	-	36.77	35.40	-
GPA	**67.60**	**67.14**	-	**66.21**	**65.81**	-	**38.14**	**37.87**	-

Table 2.3: Results on standard benchmarks, measured in P@1. * Results as reported in the original paper. **Notes**: Conneau et al. (2018) report 63.7 on Italian with Wikipedia embeddings; results with different embedding sets are not comparable due to a non-zero out-of-vocabulary rate on the test set for Wikipedia embeddings; Wikipedia embeddings are trained on corpora with removed numerals, so supervision from numerals cannot be applied.

Labaka, and Agirre (2018b). The latter introduced the second best approach to BDI known to us, VecMap, which also uses Procrustes Analysis. We conduct experiments using three forms of supervision: gold-standard seed dictionaries of 5000 word pairs, cross-lingual homographs, and numerals. We use train and test bilingual dictionaries from Dinu, Lazaridou, and Baroni (2015) for English-Italian and from Artetxe, Labaka, and Agirre (2017) for English-{Finnish, German}. Following Conneau et al. (2018), we report results with a set of CBOW embeddings trained on the WaCky corpus (Barone, 2016), and with Wikipedia embeddings.

Results are reported in Table 2.3. We observe that GPA outperforms PA consistently on Italian and German with the WaCky embeddings, and on all languages with the Wikipedia embeddings. Notice that once more, Finnish benefits the most from a switch to GPA in the Wikipedia embeddings setting, but it is also the only language to suffer from that switch in the WaCky setup.

Interestingly, PA fails to learn a good alignment for Italian and Finnish when supervised with numerals, while GPA performs comparably with numerals as with other forms of supervision. Conneau et al. (2018) point out that improvement from subsequent iterations of PA is generally negligible, which we also found to be the case. We also found that while PA learned a slightly poorer alignment than GPA, it did so faster. With our criterion for early stopping, PA converged in 5 to 10 epochs, while GPA did so within 10 to 15 epochs[7] . In the case of Italian and Finnish alignment supervised by numerals, PA converged in 8 and 5 epochs, respectively, but clearly got stuck in local minima.

[7] Notice that one epoch with both PA and GPA takes less than half a minute, so the slower convergence of GPA is in no way prohibitive.

	AF		BS		ET		HE		OC		Ave	
	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$
PA	28.87	50.53	22.40	48.40	30.00	57.93	37.53	67.27	17.12	33.26	27.18	51.48
GPA	**29.93**	**50.67**	**24.20**	**50.20**	**31.87**	**60.07**	38.93	**68.93**	17.12	34.91	28.41	52.96
MGPA	28.93	49.20	21.00	48.60	30.73	59.53	37.53	66.47	**23.82**	**40.18**	28.40	52.80
MGPA$^+$	28.80	49.20	23.46	48.87	31.27	59.80	**40.40**	68.80	22.83	38.53	**29.35**	**53.04**

Table 2.4: Results for low-resource languages with PA, GPA and two multi-support settings.

GPA took considerably longer to converge: 27 and 74 epochs, respectively, but also managed to find a reasonable alignment between the language spaces. This points to an important difference in the learning properties of PA and GPA—unlike PA, GPA has a two-fold objective of opposing forces: it is simultaneously aligning each embedding space to two others, thus pulling it in different directions. This characteristic helps GPA avoid particularly adverse local minima.

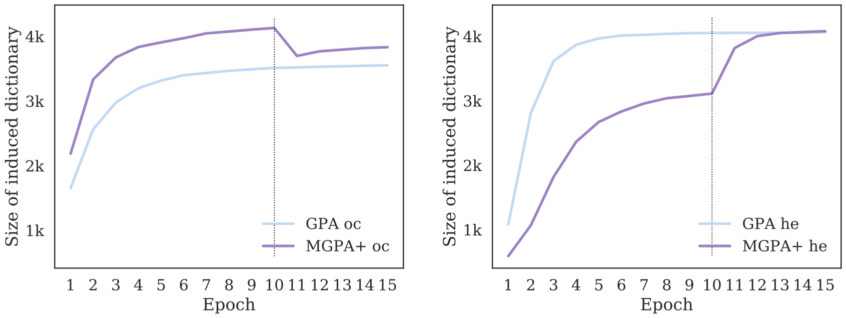
2.4.2 *Multi-support GPA*

In these experiments, we perform GPA with $k = 3$, including a third, linguistically-related supporting language in the alignment process. To best evaluate the benefits of the multi-support setup, we use as targets five low-resource languages: Afrikaans, Bosnian, Estonian, Hebrew and Occitan (see statistics in Table 2.1)[8]. Three-way dictionaries, both the initial one (consisting of cross-lingual homographs) and subsequent ones, are obtained by assuming transitivity between two-way dictionaries: if two pairs of words, e^m–e^n and e^m–e^l, are deemed translational pairs, then we consider e^n–e^m–e^l a translational triple.

We report results in Table 2.4 with multi-support GPA in two settings: a three-way alignment trained for 10 epochs (MGPA), and a three-way alignment trained for 10 epochs, followed by 5 epochs of two-way fine-tuning (MGPA+). We observe that at least one of our new methods always improves over PA. GPA always outperforms PA and it also outperforms the multi-support settings on three out of five languages. Yet, results for Hebrew and especially for Occitan, are best in a multi-support setting—we thus mostly focus on these two languages in the following subsections.

MGPA has variable performance: for four languages precision suffers from the addition of a third language, e.g. compare 38.93 for Hebrew with GPA to 37.53 with MGPA; for Occitan, however, the most challenging target language in our experiments, MGPA beats all other approaches by a large margin: 17.12 with GPA versus 23.81 with

[8] Occitan dictionaries were not available from the MUSE project, so we extracted a test dictionary of 911 unique word pairs from an English-Occitan lexicon available at `http://www.occitania.online.fr/aqui.comenca.occitania/en-oc.html`.



(a) Occitan

(b) Hebrew

Figure 2.2: Progression of dictionary size during GPA and MGPA+ training. The dotted line marks the boundary between MGPA and fine-tuning.

MGPA. This pattern relates to the effect a supporting language has on the size of the induced seed dictionary. Figure 2.2 visualizes the progression of dictionary size during training with and without a supporting language for Occitan and Hebrew. The portion of the purple curves to the left of the dotted line corresponds to MGPA: notice how the curves are swapped between the two plots. Spanish *actually* provides support for the English-Occitan alignment, by contributing to an increasingly larger seed dictionary—this provides better anchoring for the learned alignment. Having Arabic as support for English-Hebrew alignment, on the other hand, causes a considerable reduction in the size of the seed dictionaries, giving GPA less anchor points and thus damaging the learned alignment. The variable effect of a supporting language on dictionary size, and consequently on alignment precision, relates to the quality of alignment of the support language with English and with the target language: referring back to Table 2.2, English-Spanish, for example, scores at 81.93, while English-Arabic precision is 35.33. Notice that despite our linguistically-motivated choice to pair related low- and high-resource languages for multi-support training, it is not necessarily the case that those should align especially well, as that would also depend on practical factors, such as embeddings quality and training corpora similarity (Søgaard, Ruder, and Vulić, 2018).

MGPA+ applies two-way fine-tuning on top of MGPA. This leads to a drop in precision for Occitan, due to the removed support of Spanish and the consequent reduction in size of the induced dictionary (observe the fall of the purple curve after the dotted line in Figure 2.2 (a)). Meanwhile, precision for Hebrew is highest with MGPA+ out of all methods included. While Arabic itself is not a good support language, its presence in the three-way MGPA alignment seems to have resulted in a good initialization for the English-Hebrew two-way fine-tuning, thus helping the model reach an even better minimum along the loss curve.

2.5 DISCUSSION: WHY IT WORKS

If word vector spaces were completely isomorphic, the introduction of a third (or fourth) space, and the application of GPA, would lead to the same alignment as the alignment learned by PA, projecting the source language E into the target space F. This follows from the transitivity of isomorphism: if E is isomorphic to G and G is isomorphic to F, then E is isomorphic to F, via the isomorphism obtained by composing the isomorphisms from E to G and from G to F. So why do we observe improvements?

Søgaard, Ruder, and Vulić (2018) have shown that word vector spaces are often relatively far from being isomorphic, and approximate isomorphism is not transitive. What we observe therefore appears to be an instance of the Poincaré Paradox (Poincaré, 1902). While GPA is not more expressive than PA, it may still be easier to align each monolingual space to an intermediate space, as the latter constitutes a more similar target (albeit a non-isomorphic one); for example, the loss landscape of aligning a source and target language word embedding with an average of the two may be much smoother than when aligning source directly with target. Our work is in this way similar in spirit to Raiko, Valpola, and LeCun (2012), who use simple linear transforms to make learning of non-linear problems easier.

2.5.1 *Error Analysis*

Table 6.5 lists example translational pairs as induced from alignments between English and Bosnian, learned with PA, GPA and MGPA+. For interpretability, we query the system with words in Bosnian and seek their nearest neighbors in the English embedding space. P@1 over the Bosnian-English test set of Conneau et al. (2018) is 31.33, 34.80, and 34.47 for PA, GPA and MGPA+, respectively. The examples are grouped in three blocks, based on success and failure of PA and GPA alignments to retrieve a valid translation.

It appears that a lot of the difference in performance between PA and GPA concerns **morphologically related words**, e.g. *campaign* v. *campaigning*, *dialogue* v. *dialogues*, *merger* v. *merging* etc. These word pairs are naturally confusing to a BDI system, due to their related meaning and possibly identical syntactic properties (e.g. *merger* and *merging* can both be nouns). Another common mistake we observed in mismatches between PA and GPA predictions, was the wrong choice between two **antonyms**, e.g. *stable* v. *unstable* and *visible* v. *unnoticeable*. Distributional word representations are known to suffer from limitations with respect to capturing opposition of meaning

	query	gold	PA	GPA	MGPA+
PA ✗, GPA ✓	variraju	vary	varies	vary	varies
	kanjon	canyon	headwaters	canyon	headwaters
	dijalog	dialogue	dialogues	dialogue	dialogue
	izjava	statement	deniable	statement	statements
	plazme	plasma	conduction	plasma	microspheres
	računari	computers	minicomputers	computers	mainframes
	aparat	apparatus	duplex	apparatus	apparatus
	sazviježđa	constellations	asterisms	constellations	constellations
	uspostavljanje	establishing	reestablishing	establishing	establishing
	industrijska	industrial	industry	industrial	industrial
	stabilna	stable	unstable	stable	stable
	disertaciju	dissertation	habilitation	dissertation	thesis
	protivnici	opponents	opposing	opponents	opponents
	pozitivni	positive	negative	positive	positive
	instalacija	installation	installations	installation	installation
	duhana	tobacco	liquors	tobacco	tobacco
PA ✓, GPA ✗	hor	choir	choir	musicum	choir
	crijevo	intestine	intestine	intestines	intestine
	vidljiva	visible	visible	unnoticeable	visible
	temelja	foundations	foundations	superstructures	pillars
	kolonijalne	colonial	colonial	colonialists	colonialists
	spajanje	merger	merger	merging	merging
	suha	dry	dry	humid	dry
	janez	janez	janez	mariza	janez
	kampanju	campaign	campaign	campaigning	campaign
	migracije	migration	migration	migrations	migrations
	sobu	room	room	bathroom	bathroom
	predgrađu	suburb	suburb	outskirts	suburb
	specijalno	specially	specially	specialist	specially
	hiv	hiv	hiv	meningococcal	hiv
	otkrije	discover	discover	discovers	discover
	proizlazi	arises	arises	differentiates	deriving
	tajno	secretly	secretly	confidentially	secretly
PA ✗, GPA ✗	odred	squad	reconnoitre	stragglers	skirmished
	učesnik	attendee	participant	participant	participant
	saznao	learned	confided	confided	confided
	dobiva	gets	earns	earns	earns
	harris	harris	guinn	zachary	zachary
	snimke	videos	footage	footages	footage
	usne	lips	ear	ear	toes
	ukinuta	lifted	abolished	abolished	abolished
	objave	posts	publish	publish	publish
	obilježje	landmark	commemorates	commemorates	commemorates
	molim	please	appologize	thank	kindly
	čvrste	solid	concretes	concretes	concretes
	intel	intel	genesys	motorola	transputer
	transformacije	transformations	transformation	transformation	transformation

Table 2.5: Example translations from Bosnian into English.

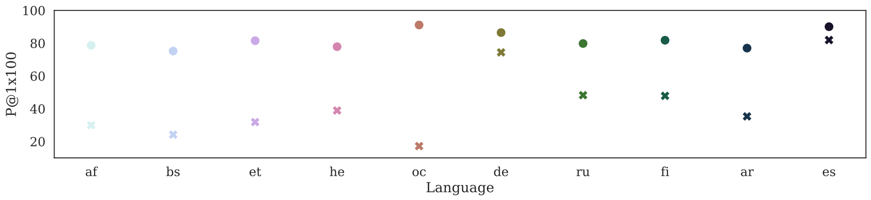


Figure 2.3: Procrustes fit test. Circles mark the results from fitting and evaluating GPA on the test dictionaries to measure the *Procrustes fit*. **x**s mark the weakly-supervised results reported in Tables 2.2 and 2.4.

(Mohammad et al., 2013), so it is not surprising that both PA- and GPA-learned alignments can fail in making this distinction. While it is not the case that GPA always outperforms PA on a query-to-query basis in these rather challenging cases, on average GPA appears to learn an alignment more robust to subtle morphological and semantic differences between neighboring words. Still, there are cases where PA and GPA both choose the wrong morphological variant of an otherwise correctly identified target word, e.g. *transformation* v. *transformations*.

Notice that many of the queries for which both algorithms fail, do result in a **nearly synonymous word** being predicted, e.g. *participant* for *attendee*, *earns* for *gets*, *footage* for *video*, etc. This serves to show that the learned alignments are generally good, but they are not sufficiently precise. This issue can have two sources: a suboptimal method for learning the alignment and/or a ceiling effect on how good of an alignment can be obtained, within the space of orthogonal linear transformations.

2.5.2 *Procrustes fit*

To explore the latter issue and to further compare the capabilities of PA and GPA, we perform a *Procrustes fit* test, where we learn alignments in a fully supervised fashion, using the test dictionaries of Conneau et al. (2018)[9] for both training *and* evaluation[10]. In the ideal case, i.e. if the subspaces defined by the words in the seed dictionaries are perfectly alignable, this setup should result in precision of 100%.

We found the difference between the fit with PA and GPA to be negligible, 0.20 on average across all 10 languages (5 low-resource and 5 high-source languages). It is not surprising that PA and GPA results in almost equivalent fits—the two algorithms both rely on linear transformations, i.e. they are equal in expressivity. As pointed out earlier, the superiority of GPA over PA stems from its more robust

[9] For Occitan, we use our own test dictionary.

[10] In this experiment, we only run a single epoch of each alignment algorithm, as that is guaranteed to give us the best Procrustes fit for the particular set of training word pairs we would then evaluate on.

learning procedure, not from higher expressivity. Figure 2.3 thus only visualizes the Procrustes fit as obtained with GPA.

The Procrustes fit of all languages is indeed lower than 100%, showing that there is a **ceiling on the linear alignability** between the source and target spaces. We attribute this ceiling effect to variable degrees of linguistic difference between source and target language and possibly to differences in the contents of cross-lingual Wikipedias (recall that the embeddings we use are trained on Wikipedia corpora). An apparent correlation emerges between the Procrustes fit and precision scores for weakly-supervised GPA, i.e. between the circles and the **x**s in the plot. The only language that does not conform here is Occitan, which has the highest Procrustes fit and the lowest GPA precision out of all languages, but this result has an important caveat: our dictionary for Occitan comes from a different source and is much smaller than all the other dictionaries.

For some of the high-resource languages, weakly-supervised GPA takes us rather close to the best possible fit: e.g. for Spanish GPA scores 81.93%, and the Procrustes fit is 90.07%. While low-resource languages do not necessarily have lower Procrustes fits than high-resource ones (compare Estonian and Finnish, for example), the gap between the Procrustes fit and GPA precision is on average much higher within low-resource languages than within high-resource ones (52.46[11] compared to 25.47, respectively). This finding is in line with the common understanding that the quality of distributional word vectors depends on the amount of data available—we can infer from these results that suboptimal embeddings results in suboptimal cross-lingual alignments.

2.5.3 *Multilinguality*

Finally, we note that there may be specific advantages to including support languages for which large monolingual corpora exist, as those should, theoretically, be easier to align with English (also a high-resource language): variance in vector directionality, as studied in Mimno and Thompson (2017), increases with corpus size, so we would expect embedding spaces learned from corpora comparable in size, to also be more similar in shape.

2.6 RELATED WORK

BILINGUAL EMBEDDINGS Many diverse cross-lingual word embedding models have been proposed (Ruder, Vulić, and Søgaard, 2018). The most popular kind learns a linear transformation from source to target language space (Mikolov, Le, and Sutskever, 2013). In most recent work, this mapping is constrained to be orthogonal

[11] Even if we leave Occitan out as an outlier, this number is still rather high: 47.10.

and solved using Procrustes Analysis (Artetxe, Labaka, and Agirre, 2017, 2018b; Conneau et al., 2018; Lu et al., 2015; Xing et al., 2015). The approach most similar to ours, Faruqui and Dyer (2014), uses canonical correlation analysis (CCA) to project both source and target language spaces into a third, joint space. In this setup, similarly to GPA, the third space is iteratively updated, such that at timestep t, it is a product of the two language spaces as transformed by the mapping learned at timestep $t-1$. The objective that drives the updates of the mapping matrices is to maximize the correlation between the projected embeddings of translational equivalents (where the latter are taken from a gold-standard seed dictionary). In their analysis of the transformed embedding spaces, Faruqui and Dyer (2014) focus on the improved quality of monolingual embedding spaces themselves and do not perform evaluation of the task of BDI. They find that the transformed monolingual spaces better encode the difference between synonyms and antonyms: in the original monolingual English space, synonyms and antonyms of *beautiful* are all mapped close to each other in a mixed fashion; in the transformed space the synonyms of *beautiful* are mapped in a cluster around the query word and its antonyms are mapped in a separate cluster. This finding is in line with our observation that GPA-learned alignments are more precise in distinguishing between synonyms and antonyms.

MULTILINGUAL EMBEDDINGS Several approaches extend existing methods to space alignments between more than two languages (Ammar et al., 2016b; Ruder, Vulić, and Søgaard, 2018). Smith et al. (2017) project all vocabularies into the English space. In some cases, multilingual training has been shown to lead to improvements over bilingually trained embedding spaces (Vulić, Mrkšić, and Korhonen, 2017), similar to our findings.

2.7 CONCLUSION

Generalized Procrustes Analysis yields benefits over simple Procrustes Analysis for Bilingual Dictionary Induction, due to its smoother loss landscape. In line with earlier research, benefits from the introduction of a common latent space seem to relate to a better distinction of synonyms and antonyms, and of syntactically-related words. GPA also offers the possibility to include multi-lingual support for inducing a larger seed dictionary during training, which better anchors the English to target language alignment in low-resource scenarios.

3

ONE STEP AT A TIME: THE IMPORTANCE OF EVALUATING ONLY THE FIRST STEP OF UNSUPERVISED WORD TRANSLATION

ABSTRACT

Cross-lingual word vector space alignment is the task of mapping the vocabularies of two languages into a shared semantic space, which can be used for dictionary induction, unsupervised machine translation, and transfer learning. In the unsupervised regime, an initial seed dictionary is learned in the absence of any known correspondences between words, through **distribution matching**, and the seed dictionary is then used to supervise the induction of the final alignment in what is typically referred to as a (possibly iterative) **refinement** step. We focus on the first step and compare distribution matching techniques in the context of language pairs for which mixed training stability and evaluation scores have been reported. We show that, surprisingly, when looking at this initial step in isolation, vanilla GANs are superior to more recent methods, both in terms of precision and robustness. The improvements reported by more recent methods thus stem from the refinement techniques, and we show that we can obtain state-of-the-art performance combining vanilla GANs with such refinement techniques.

3.1 INTRODUCTION

A word vector space – sometimes referred to as a *word embedding* – associates similar words in a vocabulary with similar vectors. Learning a projection of one word vector space into another, such that similar words – across the two word embeddings – are associated with similar vectors, is useful in many contexts, with the most prominent example being the alignment of vocabularies of different languages, i.e., word translation. This is a key step in machine translation of low-resource languages (Lample, Denoyer, and Ranzato, 2018).

Projections between word vector spaces have typically been learned from seed dictionaries. In seminal papers (Faruqui and Dyer, 2014; Gouws and Søgaard, 2015; Mikolov, Le, and Sutskever, 2013), these seeds would comprise thousands of words, but Vulić and Korhonen (2016) showed that we can learn reliable projections from as little as 50 words. Smith et al. (2017) and Hauer, Nicolai, and Kondrak (2017) subsequently showed that the seed can be replaced with just words that are identical across languages; and Artetxe, Labaka, and

Agirre (2017) showed that numerals can also do the job, in some cases; both proposals removing the need for an actual dictionary. Even more recently, entirely unsupervised approaches to projecting word vector spaces onto each other have been proposed, which induce seed dictionaries in the absence of any known correspondences between words, using distribution matching techniques. These seed dictionaries are then used as supervision for alignment algorithms based on, e.g., Procrustes Analysis (Schönemann, 1966). These unsupervised systems, in other words, typically combine two steps: an unsupervised step of distribution matching and a (possibly iterative) (pseudo-)supervised step of refinement, based on a seed dictionary learned in the first step. See Table 3.1 for an overview.

The first unsupervised dictionary induction (UBDI) systems (Barone, 2016; Conneau et al., 2018; Zhang et al., 2017) were based on Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). These approaches learn a linear transformation to minimize the divergence between a target distribution (say French word embeddings) and a source distribution (the English word embeddings projected into the French space). GAN-based approaches achieve impressive results for some language pairs (Conneau et al., 2018), but show instabilities for others. In particular, Søgaard, Ruder, and Vulić (2018) presented results suggesting that GAN-based UBDI is difficult for some language pairs exhibiting very different morphosyntactic properties, as well as when the monolingual corpora are very different. Recently, a range of unsupervised approaches that do not rely on GANs have been proposed (Artetxe, Labaka, and Agirre, 2018a; Grave, Joulin, and Berthet, 2019; Hoshen and Wolf, 2018a) in the hope they would provide a more robust alternative. In this paper, we show *none of these are more robust* on the language pairs we consider. Instead we propose a simple technique for making (vanilla) GAN-based UBDI more robust and show that combining this with a recently proposed refinement technique – stochastic dictionary induction (Artetxe, Labaka, and Agirre, 2018a) – leads to state-of-the-art performance in UBDI.

CONTRIBUTIONS We present the first systematic comparison of (a subset of) recently proposed methods for UBDI. These methods are two-step pipelines of unsupervised distribution matching for seed induction and supervised refinement. While the authors typically introduce new approaches to both steps (see Table 3.1), distribution matching and refinement are independent, and in this paper, **we focus on the distribution matching step** - by either omitting refinement or using the same refinement method across different distribution matching, or seed dictionary induction methods. On the language pairs considered here, vanilla GANs are superior to more recently improved distribution matching techniques. Moreover, we show that using an unsupervised model selection method, we can often pick out the best vanilla GAN runs *in the absence of* cross-lingual supervision. Since vanilla GANs thus seem to remain an interesting technique for inducing seed dictionaries, we explore what causes the instability of vanilla GAN seed induction, by looking at how they perform on simple transformations of the embedding spaces, and by using a combination of supervised training and model interpolation to analyze the loss landscapes. The results lead us to conclude that the instability is caused by a mild form of mode collapse, that cannot easily be overcome by changes in the number of parameters, batch size, and learning rate. Nevertheless, vanilla GANs with unsupervised model selection seem superior to more recently proposed methods, and we show that when combined with a state-of-the-art refinement technique, vanilla GANs with unsupervised model selection is superior to these methods across the board.

3.2 GAN-INITIALIZED UBDI

In this section, we discuss the dynamics of GAN-based UBDI. While the idea of using GANs for UBDI originates with Barone (2016), we refer to Conneau et al. (2018) as the canonical implementation of GAN-based UBDI. Note that GANs are not a necessary component to unsupervised distribution matchning for alignment of vector spaces, albeit a popular approach (**Conneau:ea:17**; Barone, 2016; Zhang et al., 2017). In §3, we briefly discuss how GAN-based initialization compares to the alternative of using point set registration techniques (Hoshen and Wolf, 2018a) and related strategies.

A GAN consists of a generator and a discriminator (Goodfellow et al., 2014). The generator G is trained to fool the discriminator D. The generator can be any differentiable function; in Conneau et al. (2018), it is a linear transform Ω. Let $\mathbf{e} \in E$ be an English word vector, and $\mathbf{f} \in F$ a French word vector, both of dimensionality d. The goal of the generator is then to choose $\Omega \in \mathbb{R}^{d \times d}$ such that ΩE has a distribution close to F. The discriminator is a map $D_w : \mathcal{X} \rightarrow \{0, 1\}$, implemented in Conneau et al. (2018) as a multi-layered perceptron. The objective of the discriminator is to discriminate between vector spaces F and ΩE. During training, the model parameters Ω and w are optimized using stochastic gradient descent by alternately updating the parameters of the discriminator based on the gradient of the discriminator loss and the parameters of the generator based on the gradient of the generator loss, which, by definition, is the inverse of the discriminator loss. The loss function used in Conneau et al. (2018) and in our experiments below is cross-entropy. In each iteration, we sample N vectors $e \in E$ and N vectors $f \in F$ and update the discriminator parameters w according to $w \rightarrow w + \alpha \sum_{i=1}^{N} \nabla[\log D_w(f_i) + \log(1 - D_w(G_\Omega(e_i))]$.

Theoretically, the optimal parameters are a solution to the min-max problem: $\min_\Omega \max_w \mathbb{E}[\log(D_w(F)) + \log(1 - D_w(G_\Omega(E)))]$, which reduces to $\min_\Omega JS(P_F \mid P_\Omega)$. If a generator wins the game against an ideal discriminator on a very large number of samples, then F and ΩE can be shown to be close in Jensen-Shannon divergence, and thus the model has learned the true data distribution. This result, referring to the distributions of the data, p_{data}, and the distribution, p_g, G is sampling from, is from Goodfellow et al. (2014): If G and D have enough capacity, and at each step of training, the discriminator is allowed to reach its optimum given G, and p_g is updated so as to improve the criterion $E_{\mathbf{x}\sim p_{data}}[\log D^*_G(\mathbf{x})]$ then p_g converges to p_{data}. This result relies on a number of assumptions that do not hold in practice. The generator in Conneau et al. (2018), which learns a linear transform Ω, has very limited capacity, for example, and we are updating Ω rather than p_g. In practice, therefore, during training, Conneau et al. (2018) alternate between k steps of optimizing the discriminator and one step of optimizing the generator. Another common problem with training GANs is that the discriminator loss quickly drops to zero, when there is no overlap between p_g and p_{data} (Arjovsky, Chintala, and Bottou, 2017); but note that in our case, the discriminator is initially presented with IE and F, for which there is typically no trivial solution, since the embedding spaces are likely to overlap. We show in §4 that the discriminator and generator losses are poor model selection criteria, however; instead we propose a simple criterion based on cosine similarities between nearest neighbors in the learned alignment.

From ΩE and F, a seed (bilingual) dictionary can be extracted using nearest neighbor queries, i.e., by asking for the nearest neighbor of ΩE in F, or vice versa. Conneau et al. (2018) use a normalized nearest neighbor retrieval method to reduce the influence of hubs (Dinu, Lazaridou, and Baroni, 2015; Radovanović, Nanopoulos, and Ivanovic, 2010). The method is called *cross-domain similarity local scaling* (CSLS) and used to expand high-density areas and condense low-density ones. The mean similarity of a source language embedding $\Omega\mathbf{e}$ to its k nearest neighbors in the target language is defined as $\mu^k_E(\Omega(\mathbf{e})) = \frac{1}{k}\sum_{i=1}^k \cos(\mathbf{e}, \mathbf{f}_i)$, where cos is the cosine similarity. $\mu_F(\mathbf{f}_i)$ is defined in an analogous manner for every i. $CSLS(\mathbf{e}, \mathbf{f}_i)$ is then calculated as $2\cos(\mathbf{e}, \mathbf{f}_i) - \mu_E(\Omega(\mathbf{e})) - \mu_F(\mathbf{f}_i)$. Conneau et al. (2018) use an unsupervised validation criterion based on CSLS. The translations of the top k (10,000) most frequent words in the source language are obtained with CSLS and average pairwise cosine similarity is computed over them. This metric is considered indicative of the closeness between the projected source space and the target space, and is found to correlate well with supervised evaluation metrics. After inducing a bilingual dictionary, E_d and F_d, by querying ΩE and F with CSLS, Conneau et al. (2018) perform a refinement step based on Procrustes Analysis (Schönemann, 1966). Here, the optimal mapping Ω that maps the words in the seed dictionary onto each other, is computed

	INITIALIZATION AND OPTIMIZATION STEPS		
Authors	**Unsupervised step**	**Supervised step**	**Extras**
Barone (2016)	GAN	None	
Zhang et al. (2017)	Wasserstein GAN	Procrustes	
Conneau et al. (2018)	GAN	Procrustes	
Hoshen and Wolf (2018a)	ICP	Procrustes	Restarts
Alvarez-Melis and Jaakkola (2018)	Gromov-Wasserstein	Procrustes	
Artetxe, Labaka, and Agirre (2018a)	Gromov-Wasserstein	Stochastic	
Yang et al. (2019)	Gromov-Wasserstein	MMD	
Xu et al. (2018)	GAN	Sinkhorn	Back-translation
Grave, Joulin, and Berthet (2019)	Gold-Rangarajan	Sinkhorn	

Table 3.1: Approaches to unsupervised alignment of word vector spaces. We break down these approaches in two steps (and extras): (1) **Unsupervised** distribution matching for seed dictionary learning): (W)GANs, ICP, Gromov-Wasserstein initialization, and the convex relaxation proposed in Gold and Rangarajan (1996). (2) **Supervised** refinement: Procrustes, stochastic dictionary induction, maximum mean discrepancy (MMD), and the Sinkhorn algorithm.

analytically as $\Omega = UV^T$, where U and V are obtained via the singular value decomposition $U\Sigma V^T$ of $F_d^T E_d$.

3.3 ALTERNATIVES TO GAN-INITIALIZED UBDI

This section introduces some recent alternatives to (vanilla) GAN-initialized UBDI. In Table 3.1, we list more approaches and classify them by how they perform unsupervised distribution matching and supervised refinement.

ITERATIVE CLOSEST POINT The idea of minimizing nearest neighbor distances for unsupervised model selection is also found in point set registration and lies at the core of iterative closest point (ICP) optimization (Besl and McKay, 1992). ICP typically minimizes the λ_2 distance (mean squared error) between nearest neighbor pairs. The ICP optimization algorithm works by assigning each transformed vector to its nearest neighbor and then computing the new relative transformation that minimizes the cost function with respect to this assignment. ICP can be shown to converge to local optima (Besl and McKay, 1992), in polynomial time (Ezra, Sharir, and Efrat, 2006). ICP easily gets trapped in local optima, however, exact algorithms only exist for two- and three-dimensional point set registration, and these algorithms are slow (Yang et al., 2016). Generally, it holds that the optimal solution to the GAN min-max problem is also optimal for ICP. To see this, note that a GAN minimizes the Jensen-Shannon divergence between F and ΩE. The optimal solution to this is $F = \Omega E$.

As sample size goes to infinity, this means the $\mathcal{L}_2$ loss in ICP goes to 0. In other words, the ICP loss is minimal if an optimal solution to the UBDI min-max problem is found. ICP was independently proposed for UBDI in Hoshen and Wolf (2018a). They report their method only works using PCA initialization, i.e. they project a subset of both sets of word embeddings onto the 50 first principal components, and learn an initial seed dictionary using ICP on the lower-dimensional embeddings. This seed mapping is then used as starting point for ICP on the full word embeddings. We explored PCA initialization for GAN-based distribution matching, but observed the opposite effect, namely that PCA initialization leads to a degradation in performance. The most important thing to note from Hoshen and Wolf (2018a), however, is that they do 500 random restarts of the PCA initialization to obtain robust performance; ICP, in other words, is extremely sensitive to initialization. This explains their poor performance under our experimental protocol below (Table 3.2).

WASSERSTEIN GAN Zhang et al. (2017) were the first to introduce Wasserstein GANs as a way to learn seed dictionaries in the context of UBDI. In their best system, they train simple Wasserstein GANs and use the resulting seed dictionaries to supervise Procrustes Analysis. We modified the MUSE code to experiment with Wasserstein GANs in a controlled way. Simple Wasserstein GANs were unsuccessful, but with gradient penalty (Gulrajani et al., 2017), we obtained almost competitive results, after tuning the learning rate and the gradient penalty λ using nearest neighbor cosine distance as validation criterion. On the other hand, the results were not significantly better, and instability did not improve. Finally, we experimented with CT-GANs (Wei et al., 2018), an extension of Wasserstein GANs with gradient penalty, but this only lowered performance and increased instability. Since Wasserstein GANs and CT-GANs were consistently worse and less stable than vanilla GANs, we do not include them in the experiments below.

GROMOV-WASSERSTEIN Alvarez-Melis and Jaakkola (2018) present a very different initialization strategy. In brief, Alvarez-Melis and Jaakkola (2018) learn a linear transformation to minimize Gromov-Wasserstein distances of distances between nearest neighbors, in the absence of cross-lingual supervision. We report the performance of their system in the experiments below, but results (Table 3.2) were all negative. We think the reason is that Alvarez-Melis and Jaakkola (2018) only consider small subsamples of the vector spaces, and that in hard cases, alignments induced on subspaces are unlikely to scale. It achieved an impressive P@1 of 85.6 on the Greek MUSE dataset (Conneau et al. (2018) obtain 59.5); but on the datasets, where Conneau et al. (2018) are instable, considered here, it consistently fails to align the vector spaces.

Artetxe, Labaka, and Agirre (2018a) introduce a very simple, related initialization method that is also based on Gromov-Wasserstein distances of distances between nearest neighbors: They use these second-order distances to build a seed dictionary directly by aligning nearest neighbors across languages. By itself, this is a poor initialization method (see Table 3.2). Artetxe, Labaka, and Agirre (2018a), however, combine this with a new refinement method called *stochastic dictionary induction*, i.e., randomly dropping out dimensions of the similarity matrix when extracting a seed dictionary for the next iteration of Procrustes Analysis. Artetxe, Labaka, and Agirre (2018a) show in an ablation study for one language pair (English-Finnish) that the initialization method only works in combination with the stochastic dictionary induction step, i.e., without the application of stochasticity, the induced mapping is degenerate. In our experiments below, we show that this finding generalizes to other language pairs, suggesting that the stochastic dictionary induction is the main contribution in their work. We show that when combined with vanilla GANs for the initial step of learning a seed dictionary through distribution matching, stochastic dictionary induction performs even better.

CONVEX RELAXATION The Gold-Rangarajan relaxation is a convex relaxation of the (NP-hard) graph matching problem and can be solved using the Frank-Wolfe algorithm. Once the minimal optimizer is computed, an initial transformation is obtained using singular-value decomposition. The Gold-Rangarajan relaxation can thus be used for stable learning of seed dictionaries (Grave, Joulin, and Berthet, 2019). It remains an open question how this strategy fairs on challenging language pairs such as the ones included here. We would have liked to include this approach in our experiments, but the code was not publicly available at the time of writing.

PROPERTIES OF UNSUPERVISED ALIGNMENT ALGORITHMS The above approaches provably work if the two vector spaces to be aligned, are isomorphic, except for the pathological case where the vectors are placed on an equidistant grid forming a sphere.[1] A function Ω

[1] In this case, there is an infinite set of equally good linear transformations (rotations) that achieve the same training loss. Similarly, for two binary-valued, n-dimensional vector spaces with one vector in each possible position. Here the number of local optima would be 2^n, but since the loss is the same in each of them the loss landscape is highly non-convex, and the basin of convergence is therefore very small (Yang et al., 2016). The chance of aligning the two spaces using gradient descent optimization would be $\frac{1}{2^n}$. In other words, minimizing the Jensen-Shannon divergence between the word vector distributions, even in the easy case, is not always guaranteed to uncover an alignment between translation equivalents. From the above, it follows that alignments between linearly alignable vector spaces cannot always be learned using UBDI methods. In §3.1 , we test for approximate isomorphism to decide whether two vector spaces are linearly alignable.§3.2–3.3 are devoted to analyzing *when* alignments between linearly alignable vector spaces can be learned.

from E to F is a linear transformation if $\Omega(f+g) = \Omega(f) + \Omega(g)$ and $\Omega(kf) = k\Omega(f)$ for all elements f, g of E, and for all scalars k. An invertible linear transformation is called an *isomorphism*. The two vector spaces E and F are called isomorphic, if there is an isomorphism from E to F. Equivalently, if the kernel of a linear transformation between two vector spaces of the same dimensionality contains only the zero vector, it is invertible and hence an isomorphism. Most work on supervised or unsupervised alignment of word vector spaces relies on the assumption that they are approximately isomorphic, i.e., isomorphic after removing a small set of vertices (Barone, 2016; Conneau et al., 2018; Mikolov, Le, and Sutskever, 2013; Zhang et al., 2017). It is not difficult to show that many pairs of vector spaces are not approximately isomorphic, however. See Søgaard, Ruder, and Vulić (2018) for examples.

3.4 EXPERIMENTS

In our experiments, we focus on aligning word vector spaces between two languages, by projecting from the foreign language into English. Our languages are: Estonian (et), Farsi (fa), Finnish (fi), Latvian (lv), Turkish (tr), and Vietnamese (vi). This selection of languages is motivated by observed instability when training vanilla GANs, e.g., Søgaard, Ruder, and Vulić (2018). In addition, the languages span four language families: Finno-Ugric (et, fi), Indo-European (fa, lv), Turkic (tr), and Austroasiatic (vi).

DATA In all our experiments, we use pretrained FastText embeddings (Bojanowski et al., 2017) and the bilingual test dictionaries released along with the MUSE system.[2] The FastText embeddings are trained on Wikipedia dumps[3]; the bilingual dictionaries were created using an in-house Facebook translation tool and contain translations for 1500 test words for each language pair. Since we cannot do reliable hyper-parameter optimization in the absence of cross-lingual supervision, we use MUSE with the default parameters (Conneau et al., 2018). For the experiments with stochastic dictionary induction (Table 3.3), we use the implementation in the VecMap framework (Artetxe, Labaka, and Agirre, 2018a).[4]

3.4.1 *Comparison of distribution matching strategies*

Our main experiments, reported in Table 3.2, compare the initialization strategies listed in Table 3.1: vanilla GANs, the two varieties

[2] https://github.com/facebookresearch/MUSE

[3] https://fasttext.cc/docs/en/pretrained-vectors.html

[4] https://github.com/artetxem/vecmap

		TO ENGLISH													
		et		fa		fi		lv		tr		vi		**av**	
		max	fail	max	fail	max	fail	max	fail	max	fail	max	fail	max	fail
		NO REFINEMENT													
Conneau et al.	GAN	**6.4**	9	**22.5**	3	**28.5**	1	**14.3**	9	**32.1**	2	**2.4**	9	**17.7**	5.5
Hoshen and Wolf	ICP	0.1	10	0	10	0	10	0	10	0	10	0	10	0	10
Artetxe, Labaka, and Agirre	GW	0	10	0.1	10	0.1	10	0.1	10	0.1	10	0.1	10	0.1	10
Alvarez-Melis and Jaakkola	GW	0	10	0	10	0	10	0	10	0	10	0	10	0	10
		WITH PROCRUSTES REFINEMENT													
Conneau et al.	GAN	**27.5**	9	**40.9**	3	58.9	1	**33.2**	9	**60.6**	2	**51.3**	9	**45.4**	5.5
Hoshen and Wolf	ICP	0.1	10	0	10	0	10	0	10	0	10	0	10	0	10
Artetxe, Labaka, and Agirre	GW	1.1	10	40.2	0	**60.5**	0	0.1	10	59.6	0	0.3	10	27.0	5
Alvarez-Melis and Jaakkola	GW	0	10	0	10	0	10	0	10	0	10	0	10	0	10

Table 3.2: Comparisons of unsupervised **seed dictionary** learning strategies *in the absence of refinement* (upper half) or *using the same refinement technique* (orthogonal Procrustes) (lower half). For results with refinement, we use GANs, ICPs, and Gromov-Wasserstein (GW) distribution matching and feed seed dictionaries to Procrustes refinement. We then report maximum performance (P@1) and stability (fails) across 10 runs. We consider a P@1 score below 2% a failure. The results suggest that GANs, in spite of their instability, have the highest potential for inducing useful seed dictionaries.

of Gromov-Wasserstein (see §3), and ICP.[5] Table 3.2 is split in two: First we report the performance, measured as precision at one, in the absence of refinement; and then we report the performance *with* refinement, using *the same* refinement technique (Procrustes Analysis) across the board. For all the randomly initialized algorithms (the first three), we report the best of 10 runs and the number of *fails*, where fails are runs with scores lower than 2%.[6] The reported scores are P@1, i.e., the fraction of words whose neighbors are translation equivalents.

We believe it is crucial to evaluate the different techniques this way, instead of simply comparing the numbers reported in the relevant papers: First of all, no three of these authors report performance on the same datasets. Secondly, if the authors use different refinement techniques, it is impossible to see the impact of the initialization strategies in the reported numbers. Instead we control for the refinement techniques and study the distribution matching techniques in Table 3.1 in isolation. This means, for example, that we evaluate the Artetxe, Labaka, and Agirre (2018a) in the absence of stochastic dictionary induction, and Hoshen and Wolf (2018a) in the absence of 500 random restarts. In §4.2 (Table 3.3), we compare vanilla GANs and Gromov-Wasserstein in the context of stocastic dictionary induction.

[5] We ignore Wasserstein GANs, which proved more instable than vanilla GANs in our preliminary experiments, as well as Gold-Rangarajan, which performs considerably below current state of the art.

[6] In practice, performance tends to be much higher than 2% for successful runs, hence slight changes in the threshold value would not affect results.

The patterns in Table 3.2 are very consistent. Vanilla GAN distribution matching is very instable, with 1/10 fails for Finnish and Turkish, but 6, 7 and 9 fails for Estonian, Latvian, and Vietnamese, respectively. All other methods are *more* instable, however, with the distribution matching techniques in Hoshen and Wolf (2018a) and Alvarez-Melis and Jaakkola (2018) failing across the board, with or without supervised Procrustes refinement. Vanilla GAN distribution matching also leads to higher precision for 5/6 language pairs.

Vanilla GAN distribution matching thus seems to have the highest potential for inducing useful seed dictionaries among all these methods. If we could only manage their instability, GANs seem to provide us with a better point of departure. This naturally leads us to ask: *Is it feasible to select good vanilla GAN UBDI runs from a batch of random restarts, in the absence of cross-lingual supervision?* This question is explored in §4.2, in which we also explore whether state-of-the-art performance can be achieved with vanilla GANs and a more advanced refinement technique, namely stochastic dictionary induction.

3.4.2 *GAN distribution matching with random restarts*

Exploring this question we found that the discriminator loss during training, which is used as a model selection criterion in Daskalakis et al. (2018), is a poor selection criterion. However, we did find another unsupervised model selection criterion that correlates well with UBDI performance: cosine similarity of (induced) nearest neighbors. This criterion is also used as a stopping criterion in Conneau et al. (2018), and can be used with or without CSLS scaling. This stopping criterion in fact turns out to be a quite robust model selection criterion for picking the best out of n random restarts.

In Table 3.3, we compare MUSE with 10 random restarts and using CSLS cosine similarity of nearest neighbors as an unsupervised model selection criterion, to the full state-of-the-art model in Artetxe, Labaka, and Agirre (2018a) *with* stochastic dictionary induction. What we see in these results, is that Artetxe, Labaka, and Agirre (2018a) is still superior to MUSE with random restarts, but even with 10 restarts, the gap narrows considerably, and MUSE is better on 2/6 languages. Note, however, that this is a comparison of two systems using two different refinement techniques. If we combine vanilla GAN distribution matching from MUSE with the stochastic dictionary induction technique from Artetxe, Labaka, and Agirre (2018a), we obtain slightly better performance than Artetxe, Labaka, and Agirre (2018a) (Table 3.3, mid-column): While overall improvements are small, compared to the differences in seed dictionary quality, the combination of vanilla GANs for distribution matching and stochastic dictionary induction provides a promising and fully competitive alternative to the state of the art for unsupervised word translation.

	procrustes	stochastic dictionary induction	
	C-MUSE	C-MUSE	Artetxe, Labaka, and Agirre
et-en	27.5	47.6	47.6
fa-en	40.9	**41.5**	40.2
fi-en	58.9	62.5	**63.6**
lv-en	33.2	**44.1**	41.6
tr-en	60.6	**62.8**	60.6
vi-en	51.3	**54.3**	0.3
average	45.4	**52.1**	42.3

Table 3.3: Comparison of MUSE with cosine-based model selection over 10 random restarts (C-MUSE) with and without stochastic dictionary induction (with suggested hyper-parameters from Artetxe, Labaka, and Agirre (2018a)), against state of the art. Using vanilla GANs is better than Gromov-Wasserstein on average and better on 4/6 language pairs.

3.4.3 *Discussion and Further Experiments*

We have shown that while vanilla GANs are instable, they carry a seemingly unique potential for UBDI. We have shown that a simple unsupervised cosine-based model selection criterion can achieve robust state-of-the-art performance. We have performed several other experiments to probe this instability in search of ways to stabilize vanilla GANs without significant performance drops. This subsection summarizes these experiments.

normalization We observed that GAN-based UBDI becomes more instable and performance deteriorates with unit length normalization. We performed unit length normalization (ULN) of all vectors $\mathbf{x}$, i.e., $\mathbf{x}' = \frac{\mathbf{x}}{||\mathbf{x}||^2}$, which is often used in supervised bilingual dictionary induction (Artetxe, Labaka, and Agirre, 2017; Xing et al., 2015). We used this transform to project word vectors onto a sphere – to control for shape information. If vectors are distributed smoothly over two spheres, there is no way to learn an alignment in the absence of dictionary seed; in other words, if vanilla GAN distribution matching is unaffected by this transform, vanilla GANs learn from density information alone. While supervised methods are insensitive to or benefit from ULN, we find that vanilla GANs are very sensitive to such normalization; in fact, the number of failed runs over six languages increases from below 50% to 90%. For example, while for Finnish, MUSE only fails in 1/10 runs, MUSE with ULN failed across the board; for Farsi, MUSE with ULN failed in 6/10 runs, compared to 3/10. We verify that supervised alignment is not affected by ULN by running Procrustes refinement with a seed dictionary as supervision; here, performance remains unchanged under this transformation.

NOISE INJECTION On the contrary, GAN-based UBDI is largely unaffected by noise injection. We saw this from running experiments on a few languages, but do not report performance across the board. Specifically, we add 25% random vectors, randomly sampled from a hypercube bounding the vector set. GAN-based UBDI results are not affected by noise injection. This, we found, is because the injected vectors rarely end up in the seed dictionaries used for subsequent refinement.

OVER-PARAMETERIZATION GAN training is instable because discriminators end up in poor local optima or saddle points (see below). A known technique for escaping local optima is over-parameterization (Brutzkus et al., 2018). We experimented with widening our discriminators to smoothen the loss landscape. Results were mixed, with more stability and better performance on some languages, and less stability and worse performance on others. We provide the full list of results in the Appendix.

LARGE BATCHES AND SMALL LEARNING RATES Previous work has shown that large learning rate and small batch size contribute towards SGD finding flatter minima (Jastrzebski et al., 2018), but in our experiments, we are interested in the discriminator not ending up in flat regions, where there is no signal to update the generator. We therefore experiment with (higher and) *smaller* learning rate and (smaller and) *larger* batch sizes. The motivation behind both is decreasing the scale of random fluctuations in the SGD dynamics (Balles, Romero, and Hennig, 2017; Smith and Le, 2018), enabling the discriminator to explore narrower regions in the loss landscape. Increasing the batch size or varying the learning rate (up or down), however, leads to worse performance, and it seems the MUSE default hyperparameters are close to optimal. We provide the full list of results in the Appendix.

EXPLORING THE LOSS LANDSCAPES GAN training instability arises from discriminators getting stuck in saddle points, where neither the discriminator nor the generator has a learning signals. To show this, we analyze the discriminator loss in areas of convergence by plotting it as a function of the generator parameters. Specifically, we plot the loss surface along its intersection with a line segment connecting two sets of parameters (Goodfellow, Vinyals, and Saxe, 2015; Li et al., 2018). In our case, we interpolate between the model induced by GAN-based UBDI and the (oracle) model obtained using supervised Procrustes Analysis. Results are shown in Figure 1. The green loss curves represent the current discriminator's loss along all the generators between the current generator and the generator found by Procrustes refinement. We see that while performance (P@1 and mean cosine similarity) goes up as soon as we move closer toward the

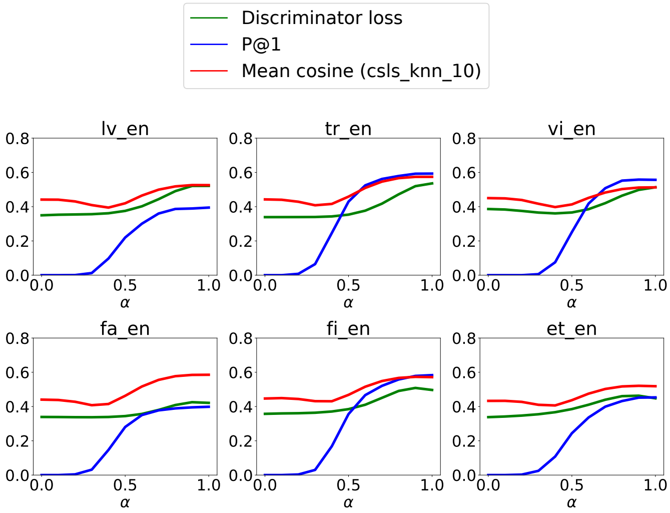


Figure 3.1: Discriminator loss averaged over all training data points (green), P@1 on the test data points (blue) and mean cosine similarity (red) on the training data – for generator parameters on the line segment that connects the unsupervised GAN solution with the supervised Procrustes Analysis solution. α is the interpolation parameter moving the generator parameters from the unsupervised GAN solution ($\alpha = 0$) to the supervised solution ($\alpha = 1$).

supervised solution, the discriminator loss does not change until we get very close to this solution, suggesting there is no learning signal in this direction for GAN-based UBDI. This is along a line segment representing the shortest path from the failed generator to the oracle generator, of course; linear interpolation provides no guarantee there are no almost-as-short paths with plenty of signal. A more sophisticated sampling method is to sample along two random direction vectors (Goodfellow, Vinyals, and Saxe, 2015; Li et al., 2018). We used an alternative strategy of sampling from normal distributions with fixed variance that were orthogonal to the line segment. We observed the same pattern, leading us to the conclusion that instability is caused by discriminator saddle points.

3.5 CONCLUSIONS

This paper explores the dynamics of (vanilla) GAN training in the context of unsupervised word translation and a systematic comparison of GANs with different distribution matching (seed induction) methods across six challenging language pairs. Our main finding is that vanilla GANs, in spite of their instability, have the highest potential for inducing useful seed dictionaries. We explore an unsupervised model selection criterion for selecting the best models from multiple random restarts, narrowing the gap between MUSE and Artetxe, Labaka, and Agirre (2018a), and further show that combining GANs with stochastic dictionary induction provides a new state of the art for unsupervised word translation.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their comments and suggestions. Mareike Hartmann was supported by the Carlsberg Foundation. Anders Søgaard was supported by a Google Focused Research Award.

4 LOST IN EVALUATION: MISLEADING BENCHMARKS FOR BILINGUAL DICTIONARY INDUCTION

4.1 ABSTRACT

The task of bilingual dictionary induction (BDI) is commonly used for intrinsic evaluation of cross-lingual word embeddings. The largest dataset for BDI was generated automatically, so its quality is dubious. We study the composition and quality of the test sets for five diverse languages from this dataset, with concerning findings: (1) a quarter of the data consists of proper nouns, which can be hardly indicative of BDI performance, and (2) there are pervasive gaps in the gold-standard targets. These issues appear to affect the ranking between cross-lingual embedding systems on individual languages, and the overall degree to which the systems differ in performance. With proper nouns removed from the data, the margin between the top two systems included in the study grows from 3.4% to 17.2%. Manual verification of the predictions, on the other hand, reveals that gaps in the gold standard targets artificially inflate the margin between the two systems on English to Bulgarian BDI from 0.1% to 6.7%. We thus suggest that future research either avoids drawing conclusions from quantitative results on this BDI dataset, or accompanies such evaluation with rigorous error analysis.

4.2 INTRODUCTION

Bilingual dictionary induction (BDI) refers to retrieving translations of individual words. The task has been widely used for intrinsic evaluation of cross-lingual embedding algorithms, which aim to map two languages into the same embedding space, for transfer learning purposes (Klementiev, Titov, and Bhattarai, 2012). Recently, Glavaš et al. (2019) reported limited evidence in support of this practice—they found that cross-lingual embeddings optimized for a BDI evaluation metric were not necessarily better on downstream tasks. Here, we study BDI evaluation in itself, as has been done for other evaluation methods in the past (cf. Faruqui et al., 2016's work on word similarity), with concerning findings about its reliability.

A massive dataset of 110 bilingual dictionaries, known as the MUSE dataset, was introduced in early 2018 along with a strong baseline (Conneau et al., 2018). Subsets of the MUSE dictionaries have been used for model comparison in the evaluation of numerous cross-lingual embedding systems developed since (cf. Grave, Joulin, and Berthet, 2019; Hoshen and Wolf, 2018a,b; Jawanpuria et al., 2019; Joulin et al., 2018a; Wada, Iwata, and Matsumoto, 2019). Even though the field has been very active, progress has been incremental for most language pairs. Moreover, there have been very few attempts at a linguistically-informed error analysis of BDI performance as measured on MUSE (cf. Kementchedjhieva et al., 2018). This is problematic for two reasons: on one hand, most systems greatly vary in their approach and architecture, so it is difficult to identify the source of the reported performance gains; on the other hand, the MUSE dataset was compiled automatically, with no manual post-processing to clean up noise, so the real impact of the performance gains is unclear.

In this work, we study the composition and quality of the MUSE data for five diverse languages: German, Danish, Bulgarian, Arabic and Hindi. A manual part-of-speech annotation of the test sets for these languages reveals a strikingly high number of proper nouns. We refer to linguistic literature to argue that proper nouns, having no lexical meaning but rather just a referential function, cannot reliably be used in the evaluation of word-level translation systems. We find that excluding proper noun pairs from the test dictionaries for the aforementioned languages affects the ranking and degree of performance gaps between five of the most influential recent systems for BDI.

With a new, more reliable ranking at hand, we perform qualitative analysis on the performance gap between the best and second best systems for Bulgarian. This reveals another major issue with the data: limited coverage of morphological variants for the target words. Through manual verification of the models' predictions, we find that the gap in performance between the two systems is far smaller than previously perceived.

The uncovered issues of high noise levels (proper nouns) and limited coverage (missing gold standard targets) clearly have a crucial impact on BDI results obtained on the MUSE dataset, and need to be addressed. Filtering out proper nouns could be achieved automatically, by checking against gazetteers of named entities. We find that an automatic procedure for the filling of missing targets, however, yields only minor improvements. We thus urge researchers to be cautious when reporting quantitative results on MUSE, and to account for the problems presented here through manual verification and analysis of the results. As an alternative, we point them to morphologically complete BDI resources, built bottom-up (Czarnowska et al., 2019). We share our part-of-speech annotations, such that future work can use this resource for analysis purposes.[1]

[1] Available at https://github.com/coastalcph/MUSE_dicos

4.3 BILINGUAL DICTIONARY INDUCTION

Improvements on BDI mostly stem from developments in the space of cross-lingual embeddings, which use BDI for intrinsic evaluation.

SYSTEMS Five influential recent systems for cross-lingual embeddings are MUSE (Conneau et al., 2018), which can be supervised (**MUSE-S**) or unsupervised (**MUSE-U**); VecMap, which also can be supervised (**VM-S**) (Artetxe, Labaka, and Agirre, 2018b) or unsupervised (**VM-U**) (Artetxe, Labaka, and Agirre, 2018a); and RCSLS (Joulin et al., 2018a), a supervised system (**RCSLS**), which scores best on BDI out of the five. We refer the reader to the respective publications for a general description of the systems.

METRICS Performance on BDI in these works is evaluated by verifying the system-retrieved translations for a source word against a set of gold-standard targets. The metric used is Precision at k (P@k), which measures how often the set of k top predictions contains one of the gold-standard targets, i.e. what is the ratio of True Positives to the sum of True Positives and False Positives.

DATA All systems listed above report results on one or both of two test sets: the MUSE test sets Conneau et al. (2018) and/or the Dinu test sets (Artetxe, Labaka, and Agirre, 2017; Dinu, Lazaridou, and Baroni, 2015). Similarly to MUSE, the Dinu dataset was compiled automatically (from Europarl word-alignments), but it only covers four languages. Due to the bigger size of MUSE (110 language pairs), we deem its impact larger and focus our study entirely on it.

4.4 ANNOTATION-BASED OBSERVATIONS

In order to gain insights into the linguistic composition of the MUSE dictionaries, we employ annotators fluent in German, Danish, Bulgarian, Arabic and Hindi (hereafter, DE, DA, BG, AR, HI) to annotate the entire dictionaries from English to one of these languages (hereafter, from- EN) and the entire dictionaries from these languages to English (hereafter, to- EN) with part-of-speech (POS) tags. Details on the annotation procedure can be found in Appendix A. Below, we discuss our findings on the POS composition of the data, and we evaluate the performance of RCSLS per POS tag.[2]

[2] For all experiments, we use the pretrained embeddings of Bojanowski et al. (2017), trained on Wikipedia.

4.4.1 *Analysis of POS composition*

The average percentage of common nouns, proper nouns, verbs, and adjectives/adverbs in the dictionaries to- EN was respectively 49.6, 24.9, 12.5, and 12.9.[3] Nouns constitute half of the dictionaries' volume, while verbs and adjectives/adverbs collectively make up only about a fourth of the average dictionary. A skewed ratio between these three categories is not surprising: in the EWT dependency treebank, for example, which contain gold-standard POS tags, the proportion of noun, verb and adjective/adverb types is 34, 17 and 14 percent, respectively. Notice, however, that in the case of the MUSE data, the ratio is even more skewed in favour of nouns over the other two categories.

The large number of proper nouns in the dictionaries seems even more problematic. Proper nouns are considered to have no lexical meaning, but rather just a referential function (Pierini, 2008). Personal names usually refer to a specific referent in a given context, but they can, in general, be attributed to different referents across different contexts, and they are almost universally interchangeable in any given context. Some personal names and most place and organization names may have a unique referent, e.g. *Barack Obama, Wisconsin, Skype*, but these names still do not carry a *sense*, their referent is resolved through access to encyclopedic knowledge (Pierini, 2008). Considering that the pretrained embeddings which we use were trained on Wikipedia, we can expect that such encyclopedic information would indeed appear in the context of certain unique names, but importantly, the alignability of the embeddings for such entities would depend on the level of parallelism between the contents of Wikipedia articles in the different languages.

With these considerations in mind, one should wonder how stable the representation of names can be in an embedding space. This question has previously been raised by Artetxe, Labaka, and Agirre (2017). We address it empirically below.

4.4.2 *Evaluation by POS*

Figure 4.1 shows the precision of the RCSLS embedding alignment method on different POS segments of the test data in mapping to- EN (results from- EN were similar and are shown in Appendix B). Verbs pose a greater challenge to BDI systems than nouns and adjectives do. Generally, we can attribute this observation to the higher abstraction of concepts described by verbs. This is a known problem for word embedding methods in general (Gerz et al., 2016), which BDI systems naturally inherit.

[3] The numbers were similar across from- EN dictionaries.

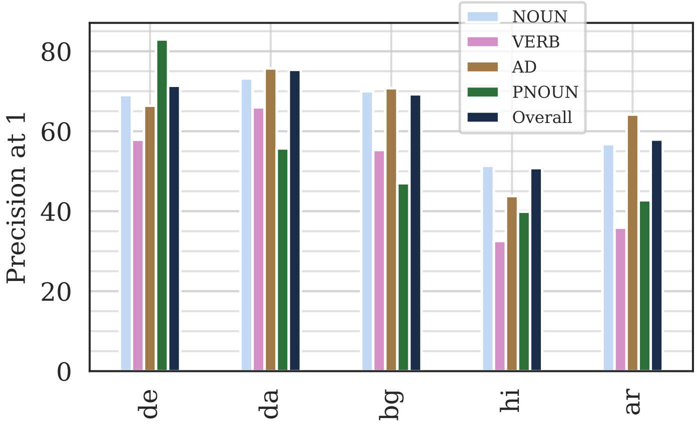


Figure 4.1: Precision of RCSLS by POS tag on to- EN data.

With respect to proper nouns, we observe that they indeed introduce a level of instability in the evaluation of BDI systems. Notice that while the other parts of speech follow a similar pattern across languages, with higher precision obtained for nouns and adjectives/adverbs than for verbs, relative precision on proper nouns is highly variable. For DE, proper nouns are easier to translate than other parts of speech by a margin of 15%, for HI and AR they are easier than nouns and adjectives/adverbs, but harder than verbs, and for DA and BG they are hardest out of all four categories. We looked into the individual word pairs marked as proper nouns in the DE and DA data, as these languages are related and RCSLS performs comparably on them otherwise, and did not find any patterns that could explain the large differences. In fact, between the 384 proper noun pairs in the EN-DE dictionary and the 330 proper noun pairs in the EN-DA dictionary, there was an overlap of 279 pairs, retrieved with precision of 89.21% in the EN-DE setting and 51.30% in the EN-DA setting. We conjecture that this result relates to the level of parallel content between the Wikipedia dumps for the different language pairs, which is likely higher for EN-DE , since the dumps for these languages are also closer in size: 5.8M articles in EN, 2.3M in DE (and only 0.2M in DA).[4]

We evaluate this hypothesis through an experiment where we train an RCSLS alignment for DE-EN using the DE embeddings of Artetxe, Labaka, and Agirre (2017), trained on SdeWaC (Baroni et al., 2009) and the EN embeddings of Dinu, Lazaridou, and Baroni (2015), trained on ukWaC (Baroni et al., 2009), Wikipedia and the BNC [5] corpora. The level of parallel content between the data used to train the two sets of embeddings is thus far more limited in this case, and the DE embeddings are not explicitly trained on Wikipedia data. Table 4.1 summarizes the results: while with the new embeddings performance is somewhat reduced for nouns, verbs and adjectives/adverbs, precision at 1 for proper nouns, in particular, drops by over 50%, indicating

[4] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[5] Available at http://www.natcorp.ox.ac.uk

Corpora	NOUN	VERB	AD	PNOUN
Wikipedia	69.0	57.9	66.4	83.0
Mixed*	64.0	55.5	59.4	37.6

Table 4.1: Comparison in performance by POS category with two different embedding sets. * The out-of-vocabulary rate for items in the dictionaries is negligible: 2, 0, and 1 for NOUN, VERB , and AD , respectively.

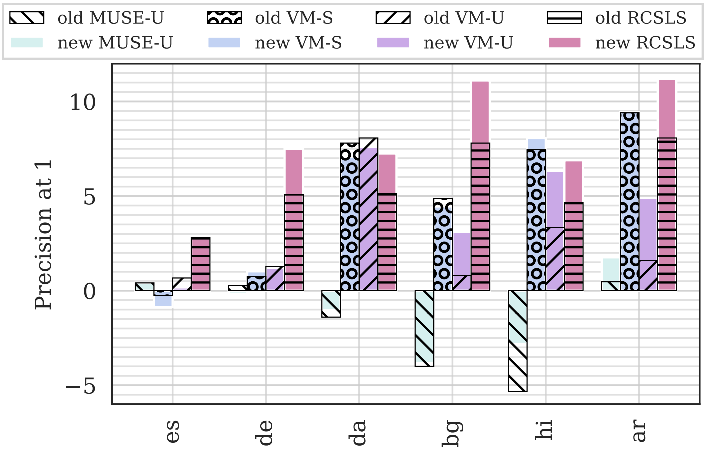


Figure 4.2: Absolute difference in performance on from- EN BDI, relative to MUSE-S. Pattern-filled bars show results as estimated on the original data (*old*), while colored bars show results as estimated on the cleaned data (*new*).

that this category of test word pairs is indeed highly sensitive to the nature of the training data.

4.4.3 *Re-ranking on clean data*

Based on the analysis presented above, we removed all pairs that were annotated as proper nouns and all pairs that were marked as invalid during the annotation process.[6] This clean-up resulted in a drop in the size of the test dictionaries of about 25% on average. A detailed size comparison between the old test dictionaries and their new cleaned versions is presented in the top rows of Table 4.4 in Appendix B. Figure 4.2 visualizes a re-evaluation of the five systems for BDI listed in Section 4.3, on the original test data and on the new clean versions of the test dictionaries from- EN.[7] The results are reported in terms of change in performance relative to MUSE-S (chosen as a baseline) as estimated on the original MUSE data (pattern-filled bars) and on the cleaned version of the data (colored bars). The absolute system

[6] The latter constitute less than 1% of the removed data.
[7] The to- EN results were similar, see Appendix B.

Ex.	SRC	TGT	RCSLS	VM-S	Description
A	joke	шега лаф виц	шега [INDEF]	шегата [DEF]	definite form missing from targets
B	remembered	запомнен	запомнен [M]	запомнена [F]	feminine form missing from targets
C	hide	скриване	скриване [N]	скриват [V]	*hide* as a verb vs. *hide* as a noun
D	bench	пейка пейката	пейка	скамейка	synonym missing from targets
E	depot	депо	депо	гара	VM-S predicted ‘train station’
F	crowned	коронован	коронована [F]	коронован [M]	feminine form missing from targets
G	pond	езерце	къщичка	езерце	RCSLS predicted ‘cottage’
H	grants	субсидии	стипендии	стипендии	synonym missing from targets
I	armies	армии	армиите	армиите	definite form missing from targets

Table 4.2: Example translations from EN to BG. Underlined forms are more canonical. Grey forms are incorrect.

performances before and after the clean-up can be found in Table 4.4 in Appendix B.

We see that the ranking between the models changes most notably for AR, where RCSLS appears inferior to VM-S on the original test data, but on the clean data it emerges as best. For BG, the evaluation on the clean test data reveals that RCSLS outperforms the next best system, VM-S, by a larger factor than it appeared on the original test data. Lastly, for DA, evaluation on the original test data makes RCSLS seem far inferior to VM-S and VM-U, but on the clean test data we see that it outperforms VM-S and matches the performance of VM-U. These observations show that the noise coming from proper nouns has a large impact on the perceived ranking and difference in performance between systems.

4.5 FALSE FALSE POSITIVES

With a more reliable estimate of the models' performance at hand, we next manually study the remaining performance gap between RCSLS, the best-performing model overall, and VM-S, the second best model overall, for EN–BG.[8] We present some examples in Table 4.2 and more can be found in Table 6.5, Appendix C.

We find that there are 125 source words that RCSLS translated correctly and VM-S did not. Upon closer inspection, we find that for 54% of these words, both RCSLS and VM-S predicted a valid translation, but RCSLS predicted a more *canonical* translation, which was listed among the gold-standard targets, while VM-S predicted another word form that was missing from the list of gold-standard targets. By more canonical we mean, for example, indefinite instead of definite forms of nouns and adjectives (see Ex. A, Table 4.2, masculine instead of feminine or neuter forms of adjectives (see Ex. B), singular instead of plural forms. To the extent that a more canonical translation

[8] We also analyzed EN– DE, with very similar results.

should be considered better, RCSLS is definitely showing superiority over VM-S. It is not clear, however, if that should be the case, since for some words, the test dictionary exhibits higher coverage than for others, i.e. the less canonical translations are not omitted by design, but appear to be accidental gaps.

Another 19% of the instances where RCSLS outperformed VM-S, we find to be clear cases of a missing translation in the test dictionary, i.e. not a missing form of a listed target, but a missing synonym or a missing sense altogether (see Ex. C and D).

The two types of errors in precision at 1 discussed above can be considered cases of *false* False Positives, because they really should have been True Positives. The remaining 27% of the gap between the two models' performance indeed illustrate that RCSLS provides better translations in some cases (see Ex. E).

Notice, however, that it is not the case that RCSLS outperformed VM-S in all cases–for 50 test words, VM-S predicted a correct translation and RCSLS did not. Among these, there are cases of missing translations from the dictionary as well (see Ex. F), but they can explain less of the lack in performance of RCSLS, i.e. 50% of the translations of RCSLS are indeed erroneous (see Ex. G).

To summarize, originally the performance gap between the two models appeared to be $(125 - 50)/1125 * 100 = 6.67\%$, while after the manual verification, it is $(27\% * 125 - 50\% * 50)/1125 * 100 = 0.1\%$.[9] Such a substantial narrowing in the gap between the two models clearly indicates that conclusions drawn on the original result, i.e. that RCSLS is far superior that VM-S for this language pair, is hardly supported by the updated result.

A surface analysis of the subset of words for which neither RCSLS nor VM-S retrieved correct translations revealed similar patterns of extensive false False Positives, due to gaps in the coverage of the dictionary (see Ex. H and I). Our takeaway from these observations is two-fold. Firstly, when RCSLS retrieves a correct target form, it also usually retrieves its most *canonical* form. More importantly, the evaluation of BDI systems on even the cleaned test dictionaries still does not represent accurately the differences in quality between them, due to major gaps in the coverage of the test dictionaries.

4.6 CONCLUDING REMARKS

Our study of the `MUSE` dataset revealed two striking problems: a high level of noise coming from proper nouns, and an issue of *false* False Positives, due to gaps in the gold-standard targets. The former problem, we conjecture, can be solved by filtering names out with gazetteers. The quality of this solution would depend on the coverage of the gazetteers. The more challenging problem, however,

[9] 1125 is the total dictionary size.

is filling in the gaps, especially in terms of inflectional forms. We carried out preliminary experiments aiming to enrich the EN–BG and EN-DE dictionaries. We extracted additional word forms of verbal and nominal targets from the UniMorph inflectional tables (Kirov et al., 2018), according to a manually designed morphosyntactic correspondence map.[10] Unfortunately, due to limited coverage of the UniMorph data, and, in the case of BG, limited vocabulary of the pretrained embeddings, the impact of this procedure was almost negligible. Alternative approaches for enrichment exists, of course, but we wonder how worthwhile further efforts would be. That is, especially in light of Glavaš et al. 2019's findings that BDI performance is not necessarily indicative of cross-lingual embedding quality. We therefore hope that our work adds weight to the call of Glavaš et al. (2019) for more reliable evaluation methods in cross-lingual embedding research. When BDI performance is used for evaluation purposes, it should be accompanied by manual verification, of the type presented here.

ACKNOWLEDGEMENTS

We thank Marcel Bollmann, Matthew Lamm, Maria Barrett, Meriem Beloucif, Mostafa Abdou, Rahul Aralikatte and Victor Petrén Hansen for help with annotations. We would also like to thank Adam Lopez, Andreas Grivas and Sameer Bansal for useful feedback on drafts of the paper, and to the anonymous reviewers for their comments and suggestions. Anders Søgaard was supported by a Google Focused Research Award; Mareike Hartmann by the Carlsberg Foundation.

[10] Details can be found in Appendix D.

A Appendix

In order to obtain a reliable part-of-speech (POS) tagging of the `MUSE` test dictionaries efficiently, we used a two-step procedure. First, we ran the Stanford POS tagger (Toutanova et al., 2003) on the English side of each dictionary. We reduced the annotation schema to five categories: nouns (NOUN), proper nouns (PNOUN), verbs (VERB), adjectives and adverbs combined (AD), and others. Next, we asked NLP researchers with the appropriate language background to verify and correct the generated tags, based on both words in a pair. Where one word in the pair is ambiguous with respect to POS, but the other is not, they were told them to use the tag of the latter. If both words were ambiguous, we told them to use the tag they considered more frequent for these words.

We instructed annotators that if a word can be both a proper noun and a common noun, it should be marked as the latter. We told them to mark pairs of identical words as proper nouns, under the assumption that they can be part of a company name or a brand, for example. That is, unless the words in the pair are actual cognates between the source and target language, or they are loanwords. See Table 4.3 for some examples. Lastly, we asked the annotators to mark pairs as invalid, if the source word is not a valid word in either the source or the target language, or the target word is not a valid translation of the source word. We note that this was a considerable annotation effort if over 40 hours in total. Each annotator had to process over 2000 word pairs: the dictionaries each consist of 1,500 source words, many of which have multiple translations, each processed separately. Annotation was performed in Microsoft Excel.

SRC	TGT	POS	valid	explanation
tea	té	NOUN	✓	actual translation
tea	tea	PNOUN	✓	part of a name, e.g. “Lipton Iced Tea”
rugby	rugby	NOUN	✓	loanword
ugby	ugby	–	✗	not a word in either language

Table 4.3: Example of annotated gold-standard word pairs from English to Spanish.

B Appendix

The pattern of performance per POS tag is similar for to- EN mappings (see Figure 4.3), as we saw it for from- EN mapping—proper nouns yield highly variable performance.

Similarly to mappings from- EN, in mappings to- EN (see Figure 4.4) we see RCSLS outperforming other systems on the clean data for all languages (and by a large margin for most of them), whereas on the original data it appeared inferior to VM-S for DA and HI. Another interesting observation here is that MUSE-U and VM-U occasionally appear inferior to the MUSE-S baseline (for DA and HI, respectively) on the original test data, but on the clean test data all models yield an improvement over the baseline.[11]

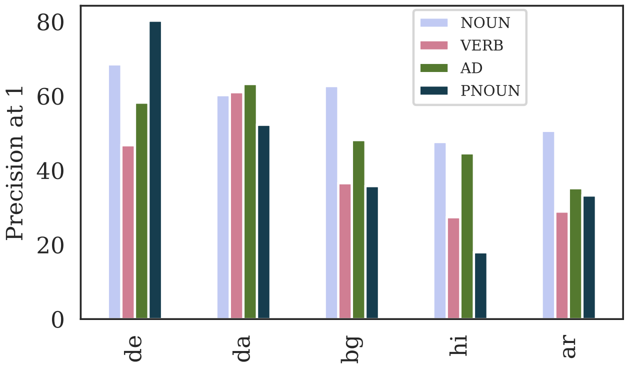


Figure 4.3: Precision of the RCSLS system, measured per POS tag, on to- EN data.

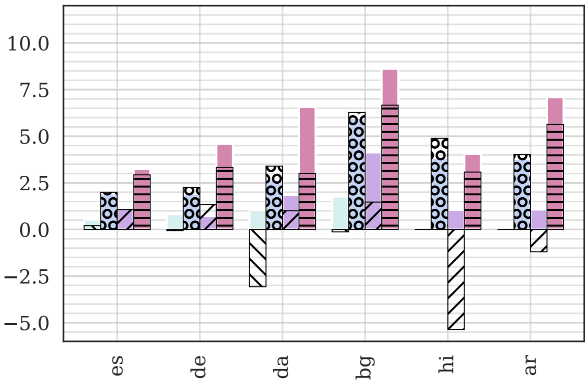


Figure 4.4: Change in performance on to- EN BDI relative to MUSE-S. Pattern-filled bars show results as estimated on the original data, while colored bars show results as estimated on the cleaned data.

[11] That is, excluding MUSE-U evaluated on HI and AR, where all solutions found were degenerate, so they have been excluded.

	es		de		da		bg		hi		ar	
	→ en	en →	→ en	en →	→ en	en →	→ en	en →	→ en	en →	→ en	en →
Source words	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500
	1145	1171	1111	1188	974	1158	1124	1125	963	1104	1212	1080
MUSE-S	83.47	81.66	72.67	73.93	67.07	56.80	56.93	43.93	44.07	33.60	49.93	34.13
	79.56	73.36	66.79	64.47	68.79	55.44	60.63	45.33	46.73	37.68	50.83	34.63
MUSE-U	83.67	82.07	72.60	74.20	64.00	55.40	56.80	39.93	0.00	28.27	0.00	34.60
	80.09	73.78	67.60	64.31	69.82	54.40	62.39	41.51	0.00	34.87	0.00	36.39
VM-S	85.47	81.40	74.93	74.67	70.47	64.60	63.20	48.80	48.96	41.07	53.95	43.53
	81.48	72.50	68.68	65.49	71.46	62.52	66.61	49.78	50.57	45.74	54.62	44.07
VM-U	84.53	82.33	74.00	75.20	68.07	64.87	58.40	44.73	38.71	36.93	48.73	35.73
	80.70	73.53	67.51	65.66	70.64	63.04	64.76	48.44	47.77	44.02	51.90	39.54
RCSLS	86.40	84.46	76.00	79.00	70.07	61.93	63.60	51.73	47.15	38.27	55.56	42.20
	82.79	76.17	71.38	71.97	75.36	62.69	69.24	56.44	50.78	44.57	57.92	45.83

Table 4.4: Cyan rows correspond to the original test data and white rows to the clean test data. The top rows report the sizes of the dictionaries, measured in terms of source words. For unstable models, e.g. MUSE-U, we train ten models and report results from one random successful model. For a fair comparison of MUSE-U and MUSE-S, we run Procrustes for 5 iterations in both cases, and use the same model selection criterion, mean cosine similarity, in both cases. All systems are evaluated using CSLS for retrieval. * Instead of full annotation for Spanish, we only mark proper nouns and remove them from the test dictionaries to and from English.

C Appendix

	SRC	TGT	RCSLS	VM-S	Description
VM-S ✗, RCSLS ✓	joke	шега лаф виц	шега	шегата	definite form missing from targets
	arbitrators	арбитри	арбитри	арбитрите	definite form missing from targets
	revolt	бунт въстание	бунт	бунта	definite form missing from targets
	remembered	запомнен	запомнен	запомнена	feminine form missing from targets
	hide	скриване	скриване	скриват	*hide* as a verb vs. *hide* as a noun
	bench	пейката пейка	пейка	скамейка	synonym missing from targets
	depot	депо	депо	гара	VM-S predicted ‘station’
	gaelic	келтски	келтски	ирландският	VM-S predicted ‘the irish’
	footage	кадри	кадри	заснети	VM-S predicted ‘shot’
VM-S ✓, RCSLS ✗	egg	яйцето яйца яйце	яйчен	яйце	translation for attributive use of noun missing from targets
	crowned	коронован	коронована	коронован	feminine form missing from targets
	volcanic	вулканична	вулканичен	вулканична	masculine form missing from targets
	penny	пени	паричка	пени	synonym missing from targets
	pound	паунд кг	кило	паунд	RCSLS predicted a non-word
	thursday	четвъртък	петък	четвъртък	RCSLS predicted ‘friday’
	striker	нападател страйкър	защитник	нападател	RCSLS predicted ‘defender’
	pond	езерце	къщичка	езерце	RCSLS predicted ‘cottage’
	flute	флейтата флейта	тромпет	флейта	RCSLS predicted ‘trumpet’
VM-S ✗, RCSLS ✗	circular	кръгло	кръгла	кръгла	feminine form missing from targets
	sailed	отплава	отплавал	отплавал	participle form missing from targets
	grants	субсидии	стипендии	стипендии	synonym missing from targets
	spots	петна	петната	петната	definite form missing from targets
	armies	армии	армиите	армиите	definite form missing from targets
	nose	нос носа носът	врат	задницата	RCSLS predicted ‘neck’, VM-S predicted ‘bottom’
	foods	храни	сладкиши	напитки	RCSLS predicted ‘sweets’, VM-S predicted ’drinks’
	cliff	скала клиф	терас	скалата	RCSLS predicted non-word, definite form missing from targets
	elevated	повишени повишена повишен	понижен	понижен	models predicted ‘reduced’

Table 4.5: Example translations from en to bg. In cases where both models predicted forms of the same word, one being more canonical than the other, we underline the canonical form. Truly incorrect translations are marked in grey. Notice the high number of correct translations that are not listed as gold-standard targets.

SRC	TGT
V;NINF	V;IMP;2;SG
	V;IMP;2;PL
	V;IND;PRS;1;SG
	V;IND;PRS;1;PL
	V;IND;PRS;2;SG
	V;IND;PRS;2;PL
	V;IND;PRS;3;PL

Table 4.6: Example of an inflectional correspondence map from English to Bulgarian.

	DE	BG
VM-S	65.5	49.8
	67.6	50.3
RCSLS	72.0	56.4
	72.5	56.8
Δ	6.5	6.7
	4.9	6.5

Table 4.7: Results before (cyan rows) and after (white rows) coverage enrichment for DE and BG.

D Appendix

Table 4.6 shows an example of an inflectional correspondence map. It signifies that whenever an English word is encountered which is a verb in the infinitive, seven Bulgarian forms would be added to the list of targets, if not in it already. Addition of targets is also conditioned on their presence in the pretrained embeddings vocabulary.

The modifications performed in this manner narrowed the gap in performance between RCSLS and VM-S by only 0.1 percentage points for EN–BG (from 6.7% to 6.4%) and by 1.6 percentage points for EN–DE (from 6.5% to 4.9%). Detailed results can be found in Table 4.7. Recall that for Bulgarian, we estimated 54% of the gap in performance to stem from false False Positives. If the enrichment procedure was perfect, it should have reduced the gap from 6.6% to less than 3.3%. Unfortunately, due to limited coverage of the inflectional tables and of the pretrained embeddings, only 240 additional word forms were added to the EN–BG dictionary, making for a an almost negligible effect on precision.

5 A SYSTEMATIC COMPARISON OF METHODS FOR LOW-RESOURCE DEPENDENCY PARSING ON GENUINELY LOW-RESOURCE LANGUAGES

ABSTRACT

Parsers are available for only a handful of the world's languages, since they require lots of training data. How far can we get with just a small amount of training data? We systematically compare a set of simple strategies for improving low-resource parsers: data augmentation, which has not been tested before; cross-lingual training; and transliteration. Experimenting on three typologically diverse low-resource languages—North Sámi, Galician, and Kazakh—we find that (1) when only the low-resource treebank is available, data augmentation is very helpful; (2) when a related high-resource treebank is available, cross-lingual training is helpful and complements data augmentation; and (3) when the high-resource treebank uses a different writing system, transliteration into a shared orthographic spaces is also very helpful.

5.1 INTRODUCTION

Large annotated treebanks are available for only a tiny fraction of the world's languages, and there is a wealth of literature on strategies for parsing with few resources (Hwa et al., 2005; McDonald, Petrov, and Hall, 2011; Søgaard, 2011; Zeman and Resnik, 2008). A popular approach is to train a parser on a related high-resource language and adapt it to the low-resource language. This approach benefits from the availability of Universal Dependencies (UD; Nivre et al., 2016), prompting substantial research (Agić, 2017; Rosa and Mareček, 2018; Tiedemann and Agic, 2016), along with the VarDial and the CoNLL UD shared tasks (Zampieri et al., 2017; Zeman et al., 2018, 2017).

But low-resource parsing is still difficult. The organizers of the CoNLL 2018 UD shared task (Zeman et al., 2018) report that, in general, results on the task's nine low-resource treebanks "are extremely low and the outputs are hardly useful for downstream applications." So if we want to build a parser in a language with few resources, what can we do? To answer this question, we systematically compare several practical strategies for low-resource parsing, asking:

1. What can we do with only a very small *target* treebank for a low-resource language?

2. What can we do if we also have a *source* treebank for a related high-resource language?

3. What if the source and target treebanks do not share a writing system?

Each of these scenarios requires different approaches. **Data augmentation** is applicable in all scenarios, and has proven useful for low-resource NLP in general (Bergmanis et al., 2017; Fadaee, Bisazza, and Monz, 2017; Sahin and Steedman, 2018). Transfer learning via **cross-lingual training** is applicable in scenarios 2 and 3. Finally, **transliteration** may be useful in scenario 3.

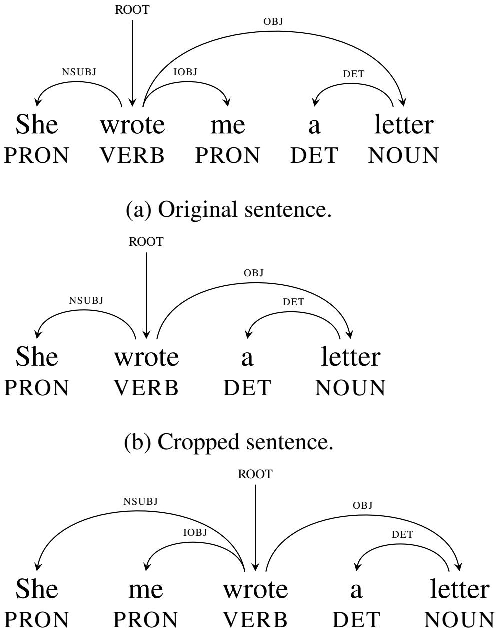
To keep our scenarios as realistic as possible, we assume that no taggers are available since this would entail substantial annotation. Therefore, our neural parsing models must learn to parse from words or characters—that is, they must be **lexicalized**—even though there may be little shared vocabulary between source and target treebanks. While this may intuitively seem to make cross-lingual training difficult, recent results have shown that lexical parameter sharing on characters and words can in fact improve cross-lingual parsing (Lhoneux et al., 2018); and that in some circumstances, a lexicalized parser can outperform a delexicalized one, even in a low-resource setting (Falenska and Çetinoğlu, 2017).

We experiment on three language pairs from different language families, in which the first of each is a genuinely low-resource language: North Sámi and Finnish (Uralic); Galician and Portuguese (Romance); and Kazakh and Turkish (Turkic), which have different writing systems[1]. To avoid optimistic evaluation, we extensively experiment only with North Sámi, which we also analyse to understand *why* our cross-lingual training outperforms the other parsing strategies. We treat Galician and Kazakh as truly held-out, and test only our best methods on these languages. Our results show that:

1. When no source treebank is available, data augmentation is very helpful: dependency tree morphing improves labeled attachment score (LAS) by as much as 9.3%. Our analysis suggests that syntactic rather than lexical variation is most useful for data augmentation.

2. When a source treebank is available, cross-lingual parsing improves LAS up to 16.2%, but data augmentation still helps, by an additional 2.6%. Our analysis suggests that improvements from cross-lingual parsing occur because the parser learns syntactic regularities about word order, since it does not have access to POS and has little reusable information about word forms.

[1] We select high-resource language based on language family, since it is the most straightforward way to define language relatedness. However, other measurement (e.g., WALS (Dryer and Haspelmath, 2013) properties) might be used.



(a) Original sentence.

(b) Cropped sentence.

(c) Rotated sentence.

Figure 5.1: Examples of dependency tree morphing operations on the sentence "She wrote me a letter".

3. If source and target treebanks have different writing systems, transliterating them to a common orthography is very effective.

5.2 methods

We describe three techniques for improving low-resource parsing: (1) two data augmentation methods which have not been applied before for dependency parsing, (2) cross-lingual training, and (3) transliteration.

5.2.1 *Data augmentation by dependency tree morphing (Morph)*

Sahin and Steedman (2018) introduce two operations to augment a dataset for low-resource POS tagging. Their method assumes access to a dependency tree, but they do not test it for dependency parsing, which we do here for the first time. The first operation, *cropping*, removes some parts of a sentence to create a smaller or simpler, meaningful sentence. The second operation, *rotation*, keeps all the words in the sentence but re-orders subtrees attached to the *root* verb, in particular those attached by nsubj (nominal subject), obj (direct object), iobj (indirect object), or obl (oblique nominal) dependencies. Figure 5.1 illustrates both operations.

It is important to note that while both operations change the set of words or the word order, they do not change the dependencies. The sentences themselves may be awkward or ill-formed, but the corresponding analyses are still likely to be correct, and thus beneficial for learning. This is because they provide the model with more examples of variations in argument structure (cropping) and in constituent order (rotation), which may benefit languages with flexible word order and rich morphology. Some of our low-resource languages have these properties—while North Sámi has a fixed word order (SVO), Galician and Kazakh have a relatively free word order. All three languages use case marking on nouns, so word order may not be as important for correct attachment.

Both rotation and cropping can produce many trees. We use the default parameters given in (Sahin and Steedman, 2018).

5.2.2 *Data augmentation by nonce sentence generation (Nonce)*

Our next data augmentation method is adapted from Gulordava et al. (2018). The main idea is to create *nonce* sentences by replacing some of the words which have the same syntactic annotations. For each training sentence, we replace each content word—nouns, verbs, or adjectives—with an alternative word having the same universal POS, morphological features, and dependency label.[2] Specifically, for each content word, we first stochastically choose whether to replace it; then, if we have chosen to replace it, we uniformly sample the replacement word type meeting the corresponding constraints. For instance, given a sentence "*He borrowed a book from the library.*", we can generate the following sentences:

1. He <u>bought</u> a book from the <u>shop</u> .
2. He <u>wore</u> a <u>umbrella</u> from the library .

This generation method is only based on syntactic features (i.e., morphology and dependency labels), so it sometimes produces nonsensical sentences like 2. But since we only replace words if they have the same morphological features and dependency label, this method preserves the original tree structures in the treebank. Following (Gulordava et al., 2018), we generate five nonce sentences for each original sentence.

5.2.3 *Cross-lingual training*

When a source treebank is available, model transfer is a viable option. We perform model transfer by cross-lingual parser training: we first train on both source and target treebanks to produce a single model, and then fine tune the model only on the target treebank. In our preliminary experiments (Appendix A), we found that fine tuning

[2] The dependency label constraint is new to this paper.

on the target treebank was effective in all settings, so we use it in all applicable experiments reported in this paper.

5.2.4 *Transliteration*

Two related languages might not share a writing system even when they belong to the same family. We evaluate whether a simple transliteration would be helpful for cross-lingual training in this case. In our study, the Turkish treebank is written in extended Latin while the Kazakh treebank is written in Cyrillic. This difference potentially makes model transfer less useful, and means we might not be able to leverage lexical similarities between the two languages. We preprocess both treebanks by transliterating them to the same "pivot" alphabet, basic Latin.[3]

The mapping from Turkish is straightforward. Its alphabet consists of 29 letters, 23 of which are in basic Latin. The other six letters, 'ç', 'ğ', 'ı', 'ö', 'ş', and 'ü', add diacritics to basic Latin characters, facilitating different pronunciations.[4] We map these to their basic Latin counterparts, e.g., 'ç' to 'c'. For Kazakh, we use a simple dictionary created by a Kazakh computational linguist to map each Cyrillic letter to the basic Latin alphabet.[5].

5.3 EXPERIMENTAL SETUP

5.3.1 *Dependency Parsing Model*

We use the Uppsala parser, a transition-based neural dependency parser (Kiperwasser and Goldberg, 2016; Lhoneux et al., 2017; Lhoneux, Stymne, and Nivre, 2017). The parser uses an arc-hybrid transition system (Kuhlmann, Gómez-Rodrìguez, and Satta, 2011), extended with a static-dynamic oracle and SWAP transition to allow non-projective dependency trees (Nivre, 2009).

Let $w = w_0, \ldots, w_{|w|}$ be an input sentence of length $|w|$ and let w_0 represent an artificial ROOT token. We create a vector representation for each input token w_i by concatenating $(;)$ its word embedding, $\mathbf{e}_w(w_i)$ and its character-based word embedding, $\mathbf{e}_c(w_i)$:

$$\mathbf{x}_i = [\mathbf{e}_w(w_i); \mathbf{e}_c(w_i)] \tag{5.1}$$

Here, $\mathbf{e}_c(w_i)$ is the output of a *character-level* bidirectional LSTM (biLSTM) encoder run over the characters of w_i (Ling et al., 2015); this

[3] Another possible pivot is phonemes (Tsvetkov et al., 2016). We leave this for future work.

[4] https://www.omniglot.com/writing/turkish.htm

[5] The mapping from Kazakh Cyrilic into basic Latin alphabet is provided in Appendix B

makes the model fully open-vocabulary, since it can produce representations for any character sequence. We then obtain a *context-sensitive* encoding $\mathbf{h}_i$ using a *word-level* biLSTM encoder:

$$\mathbf{h}_i = [\text{LSTM}_f(\mathbf{x}_{0:i}); \text{LSTM}_b(\mathbf{x}_{|w|:i})] \tag{5.2}$$

We then create a configuration by concatenating the encoding of a fixed number of words on the top of the stack and the beginning of the buffer. Given this configuration, we predict a transition and its arc label using a multi-layer perceptron (MLP). More details of the core parser can be found in (Lhoneux et al., 2017; Lhoneux, Stymne, and Nivre, 2017).

5.3.2 *Parameter sharing*

To train cross-lingual models, we use the strategy of Lhoneux et al. (2018) for parameter sharing, which uses *soft* sharing for word and character parameters, and *hard* sharing for the MLP parameters. Soft parameter sharing uses a language embedding, which, in theory, learns what parameters to share between the two languages. Let $\mathbf{c}_j$ be an embedding of character c_j in a token w_i from the treebank of language k, and let $\mathbf{l}_k$ be the language embedding. For sharing on characters, we concatenate character and language embedding: $[\mathbf{c}_j; \mathbf{l}_k]$ for input to the character-level biLSTM. Similarly, for input to the word-level biLSTM, we concatenate the language embedding to the word embedding, modifying Eq. 5.1 to

$$\mathbf{x}_i = [\mathbf{e}_w(w_i); \mathbf{e}_c(w_i); \mathbf{l}_k] \tag{5.3}$$

We use the default hyperparameters of Lhoneux et al. (2018) in our experiments. We fine-tune each model by training it further only on the target treebank (Shi, Padhi, and Knight, 2016). We use early stopping based on Label Attachment Score (LAS) on a development set.

5.3.3 *Datasets*

We use Universal Dependencies (UD) treebanks version 2.2 (Nivre et al., 2018). Our target treebanks are North Sámi Giella (Sheyanova and Tyers, 2017), Galician TreeGal, and Kazakh KTB (Makazhanov et al., 2015; Tyers and Washington, 2015). None of these treebanks have a development set, so we generate new train/dev splits by 50:50. Having large development sets allows us to perform better analysis for this study. We use Finish TDT, Portuguese Bosque (Rademaker et al., 2017), and Turkish IMST for our source treebanks. Table 5.1 shows the statistics of our datasets.

Language	Treebank ID	train	dev.	test
Finnish	fi_tdt	14981	1875	1555
North Sámi	sme_giella	1128	1129	865
Portuguese	pt_bosque	8329	560	477
Galician	gl_treegal	300	300	400
Turkish	tr_imst	3685	975	975
Kazakh	kk_ktb	15	16	1047

Table 5.1: Train/dev split used for each treebank.

	original	+Morph	+Nonce
$\mathcal{T}_{100}$	1128	7636	4934
$\mathcal{T}_{50}$	564	3838	2700
$\mathcal{T}_{10}$	141	854	661

Table 5.2: Number of North Sámi training sentences.

5.4 parsing north sámi

North Sámi is our largest low-resource treebank, so we use it for a full evaluation and analysis of different strategies before testing on the other languages. To understand the effect of target treebank size, we generate three datasets with different *training* sizes: $\mathcal{T}_{10}$ (~10%), $\mathcal{T}_{50}$ (~50%), and $\mathcal{T}_{100}$ (100%). Table 5.2 reports the number of training sentences after we augment the data using the methods described in Section 5.2. We apply Morph and Nonce separately to understand the effect of each method and to control the amount of noise in the augmented data.

We employ two baselines: a monolingual model (§5.3.1) and a cross-lingual model (§5.2.3), both *without* data augmentation. The monolingual model acts as a simple baseline, to resemble a situation when the target treebank does not have any source treebank (i.e., no available treebanks from related languages). The cross-lingual model serves as a strong baseline, simulating a case when there is a source treebank. We compare both baselines to models trained with Morph and Nonce augmentation methods. Table 5.3 reports our results, and we review our motivating scenarios below.

scenario 1: we only have a very small target treebank. In the monolingual experiments, we observe that both dependency tree morphing (Morph) and nonce sentence generation (Nonce) improve performance, indicating the strong benefits of data augmentation when there are no other resources available except the target treebank

	MONOLINGUAL			CROSS-LINGUAL		
size	mono-base	+Morph	+Nonce	cross-base	+Morph	+Nonce
$\mathcal{T}_{100}$	53.3	56.0 (+3.3)	56.3 (+3.0)	61.3 (+8.0)	60.9 (+7.6)	**61.7 (+8.4)**
$\mathcal{T}_{50}$	42.5	46.6 (+4.1)	46.5 (+4.0)	52.0 (+9.5)	51.7 (+9.2)	**52.0 (+9.5)**
$\mathcal{T}_{10}$	18.5	27.1 (+8.6)	27.8 (+9.3)	34.7 (+16.2)	**37.3 (+18.8)**	35.4 (+16.9)

Table 5.3: LAS results on North Sámi development data. *mono-base* and *cross-base* are models without data augmentation. % improvements over *mono-base* shown in parentheses.

itself. In particular, when the number of training data is the lowest ($\mathcal{T}_{10}$), data augmentations improves performance up to 9.3% LAS.

SCENARIO 2: A SOURCE TREEBANK IS AVAILABLE. We see that the cross-lingual training (cross-base) performs better than monolingual models even with augmentation. For the $\mathcal{T}_{10}$ setting, cross-base achieves almost twice as much as the monolingual baseline (mono-base). The benefits of data augmentation are less evident in the cross-lingual setting, but in the $\mathcal{T}_{10}$ scenario, data augmentation still clearly helps. Overall, cross-lingual combined with data augmentation yields the best result.

5.4.1 *What is learned from Finnish?*

Why do cross-lingual training and data augmentation help? To put this question in context, we first consider their relationship. Finnish and North Sámi are mutually unintelligible, but they are typologically similar: of the 49 (mostly syntactic) linguistic features annotated for North Sámi in the Word Atlas of Languages (WALS; Dryer and Haspelmath, 2013), Finnish shares the same values for 42 of them.[6] Despite this and their phylogenetic and geographical relatedness, they share very little vocabulary: only 6.5% of North Sámi tokens appear in Finnish data, and these words are either proper nouns or closed class words such as pronouns or conjunctions. However, both languages do share many character-trigrams (72.5%, token-level), especially in terms of suffixes.

Now we turn to an analysis of the $\mathcal{T}_{10}$ data setting, where we see the largest gains for all methods.

5.4.2 *Analysis of data augmentation*

For dependency parsing, POS features are important because they can provide a strong signal as to whether there exists dependency

[6] There are 192 linguistic features in WALS, but only 49 are defined for North Sámi. These features are mostly syntactic, annotated within different areas such as morphology, phonology, nominal and verbal categories, and word order.

POS	%dev	baseline	%diff. with	
			+Morph	+Nonce
INTJ	0.1	0.0	20.0	20.0
PART	1.5	70.1	7.7	0.8
NUM	1.9	19.2	15.1	-4.1
ADP	1.9	15.7	24.5	19.7
SCONJ	2.4	57.8	5.9	7.6
AUX	3.2	26.3	27.2	-4.9
CCONJ	3.4	91.3	-0.8	-4.2
PROPN	4.7	5.9	5.9	-5.9
ADJ	6.5	12.7	3.8	0.2
ADV	9.0	42.9	11.8	11.5
PRON	13.4	63.2	5.4	-2.7
VERB	25.7	72.4	-6.2	-4.5
NOUN	26.4	67.0	8.6	13.2

Table 5.4: Results for the monolingual POS predictions, ordered by the frequency of each tag in the dev split (%dev). %diff shows the difference between each augmentation method and monolingual models.

between two words in a given sentence. For example, *subject* and *object* dependencies often occur between a NOUN and a VERB, as can be seen in Fig. 5.1a. We investigate the extent to which data augmentation is useful for learning POS features, using diagnostic classifiers (Adi et al., 2017; Shi, Padhi, and Knight, 2016; Veldhoen, Hupkes, and Zuidema, 2016) to probe our model representations. Our central question is: do the models learn useful representations of POS, despite having no direct access to it? And if so, is this helped by data augmentation?

After training each model, we freeze the parameters and generate *context-dependent* representations (i.e., the output of *word-level* biLSTM, $\mathbf{h}_i$ in Eq. 5.2), for the training and development data. We then train a feed-forward neural network classifier to predict the POS tag of each word, using only the representation as input. To filter out the effect of cross-lingual training, we only analyze representations trained using the *monolingual* models. Our training and development data consists of 6321 and 7710 tokens, respectively. The percentage of OOV tokens is 40.5%.

Table 5.4 reports the POS prediction accuracy. We observe that representations generated with monolingual Morph seem to learn better POS, for most of the tags. On the other hand, representations generated with monolingual Nonce sometimes produce lower accuracy on some tags; only on nouns the accuracy is better than monolingual Morph. We hypothesize that this is because Nonce sometimes generates meaningless sentences which confuse the model. In parsing this effect is less apparent, mainly because monolingual Nonce has the poorest POS representation for infrequent tags (%dev), and better representation of nouns.

5.4.3 *Effects of cross-lingual training*

Next, we analyze the effect of cross-lingual training by comparing the monolingual baseline to the cross-lingual model with Morph.

cross-lingual representations. The fact that the cross-lingual model improves parsing performance is interesting, since Finnish and North Sámi have so little common vocabulary. What linguistic knowledge is transferred through cross-lingual training? We analyze whether words with the same POS category from the source and target treebanks have similar representations. To do this, we analyze the *head predictions*, and collect North Sámi tokens for which only the cross-lingual model correctly predicts the headword.[7] For these words, we compare token-level representations of North Sámi *development* data to Finnish *training* data.

We ask the following questions: Given the representation of a North Sámi word, what is the Finnish word with the most similar representation? Do they share the same POS category? Information other than POS may very well be captured, but we expect that the representations will reflect similar POS since POS is highly revelant to parsing. We use *cosine distance* to measure similarity.

We look at four categories for which cross-lingual training substantially improves results on the development set: adjectives, nouns, pronouns, and verbs. We analyze the representations generated by two layers of the model in §5.3.1: (1) the output of character-level biLSTM (char-level), $\mathbf{e}_c(w_i)$ and (2) the output of word-level biLSTM (word-level), i.e., $\mathbf{h}_i$ in Eq. 5.2.

Table 5.5 shows examples of the top three closest Finnish training words for a given North Sámi word. We observe that the character-level representation focuses on orthographic similarity of suffixes, rather than POS. On the level of word representations, we find more cases when the top closest Finnish words have the same POS with the North Sámi word. In fact, when we compare the most similar Finnish word (Table 5.6) quantitatively, we find that the word-level representations of North Sámi are often similar to Finnish word with the same POS; the same trend does not hold for character-level representations. Since very few word tokens are shared, this suggests that improvements in cross-lingual training might simply be due to syntactic (i.e. word order) similarities between the two languages,

[7] Another possible way is to look at the label predictions. But since the monolingual baseline LAS is very low, we focus on the unlabeled attachment prediction since it is more accurate.

	Top nearest Finnish words	
North Sámi	char-level	word-level
borrat	*herrat* (NOUN; gentleman)	*käydä* (VERB; go)
(VERB; eat)	*kerrat* (NOUN; time)	*otan* (VERB; take)
	naurat (VERB; laugh)	*sain* (VERB; get)
veahki	*nuuhki* (VERB; sniff)	*tyhjäksi* (ADJ; empty)
(NOUN; help)	*väki* (NOUN; power)	*johonki* (PRON; something)
	avarsi (VERB; expand)	*lähtökohdaksi* (NOUN; basis)
divrras	*harras* (ADJ; devout)	*välttämätöntä* (ADJ; essential)
(ADJ; expensive)	*reipas* (ADJ; brave)	*mahdollista* (ADJ; possible)
	sarjaporras (NOUN; series)	*kilpailukykyisempi* (ADJ; competitive)

Table 5.5: Most similar Finnish words for each North Sámi word based on cosine similarity.

POS	char-level (%)	word-level (%)
ADJ	12.1	37.1
NOUN	55.8	63.5
PRON	12.9	68.0
VERB	34.2	69.0

Table 5.6: Number of North Sámi tokens for which the most similar Finnish word has the same POS.

captured in the dynamics of the biLSTM encoder—despite the fact that it knows very little about the North Sámi tokens themselves. The word-level representation has an advantage over the character-level representation in that it has access to contextual information like word order, and it has knowledge about the other words in the sentence.

HEAD AND LABEL PREDICTION. Lastly, we analyze the parsing performance of the monolingual compared to the cross-lingual models. Looking at the produced parse trees, one striking difference is that the monolingual model sometimes predicts a “rootless" tree. That is, it fails to assign a head with index ‘0’ to any word and to label the dependency with a *root* label. In cases where the monolingual model predicts wrong parses and the cross-lingual model predicts the correct ones, we find that the “rootless" trees are predicted more than 50% of the time.[8] Meanwhile, the cross-lingual model learns to assign a word with head index ‘0’, although sometimes it is the incorrect word (e.g., it is the second word, but the parser predicts the fifth word). This

[8] The parsing model enforces the constraint that every tree should have a head, i.e., an arc pointing from a dummy root to a node in the tree. It does not, however, enforce that this arc be labeled *root*—the model must learn the labeling.

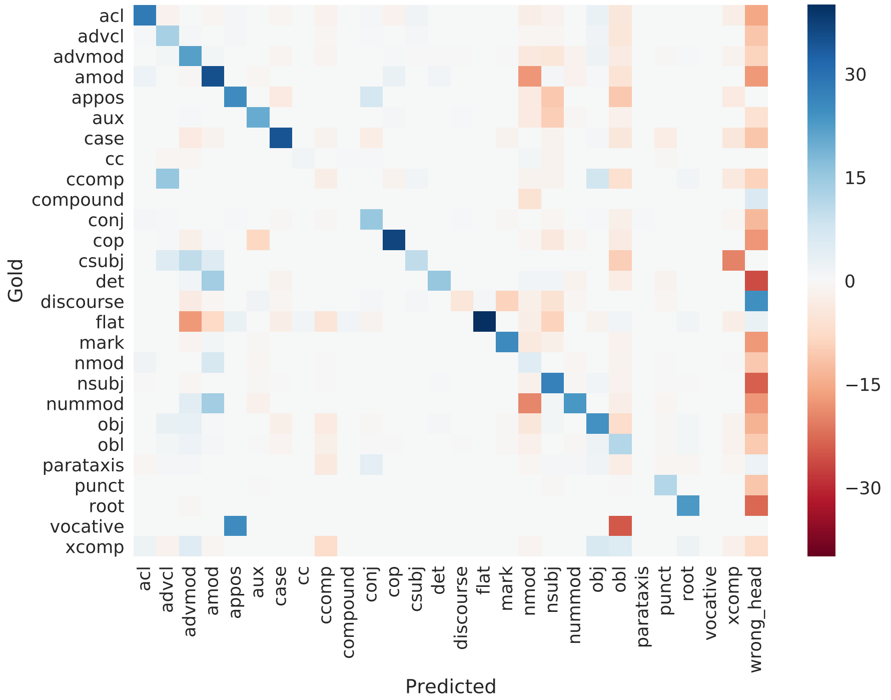


Figure 5.2: Differences between cross-lingual vs. monolingual confusion matrices. The last column represents cases of *incorrect* heads and the other columns represent cases for *correct* heads, i.e., each row summing to 100%. Blue cells show higher cross-lingual values and red cells show higher monolingual values.

pattern suggests that more training examples at least helps the model to learn the structural properties of a well-formed tree.

The ability of a parser to predict labels is contingent on its ability to predict heads, so we focus our analysis on two cases. How do monolingual and cross-lingual head prediction compare? And if both models predict the correct head, how do they compare on label prediction?

Figure 5.2 shows the *difference* between two confusion matrices: one for cross-lingual and one for monolingual models. The last column shows cases of *incorrect* heads and the other columns show label predictions when the heads are *correct*, i.e., each row sums to 100%. Here, blue cells highlight confusions that are more common for the cross-lingual model, while red cells highlight those more common for the monolingual model. For head prediction (last column), we observe that the monolingual model makes higher errors especially for nominals and modifier words. In cases when both models predict the correct heads, we observe that cross-lingual training gives further improvements in predicting most of the labels. In particular, regarding the “rootless" trees discussed before, we see evidence that cross-lingual training helps in predicting the correct root index, and the correct *root* label.

5.5 PARSING TRULY LOW-RESOURCE LANGUAGES

Now we turn to two truly low-resource treebanks: Galician and Kazakh. These treebanks are most analogous to the North Sámi $\mathcal{T}_{10}$ setting and therefore we apply the best approach, cross-lingual training with MORPH augmentation. Table 5.1 provides the statistics of the augmented data. For Galician, we use the Portuguese treebank as source while for Kazakh we use Turkish. Portuguese and Galician have high vocabulary overlap; 62.9% of Galician tokens appear in Portuguese data, while for Turkish and Kazakh they do not share vocabulary since they use different writing systems. However, after transliterating them into the same basic Latin alphabet, we observe that 9.5% of Kazakh tokens appear in the Turkish data. Both language pairs also share many (token-level) character trigrams: 96.0% for Galician-Portuguese and 66.3% for transliterated Kazakh-Turkish.

To compare our best approach, we create two baselines: (1) a pre-trained parsing model of the source treebank (zero-shot learning), and (2) a cross-lingual model initialized with *monolingual* pre-trained word embeddings. The first serves as a weak baseline, in a case where training on the target treebank is not possible (e.g., Kazakh only has 15 sentences for training). The latter serves as a strong baseline, in a case when we have access to pre-trained word embeddings, for the source and/or the target languages.

We treat a pre-trained word embedding as an external embedding, and concatenate it with the other representations, i.e., modifying Eq. 5.3 to $\mathbf{x}_i = [\mathbf{e}_w(w_i); \mathbf{e}_p(w_i); \mathbf{e}_c(w_i); \mathbf{l}_k]$, where $\mathbf{e}_p(w_i)$ represents a pre-trained word embedding of w_i, which we update during training. We use the pre-trained monolingual fastText embeddings (Bojanowski et al., 2017).[9] We concatenate the source and target pre-trained word embeddings.[10] For our experiments with transliteration (§5.2.4), we transliterate the entries of both the source and the target pre-trained word embeddings.

5.5.1 *Experimental results*

Table 5.7 reports the LAS performance on the development sets. MORPH augmentation improves performance over the zero-shot baseline and achieves comparable or better LAS with a cross-lingual model trained with pre-trained word embeddings.

Next, we look at the effects of transliteration (see Kazakh vs Kazakh (translit.) in Table 5.7). In the zero-shot experiments, simply mapping both Turkish and Kazakh characters to the Latin alphabet improves

[9] The embeddings are available at https://fasttext.cc/docs/en/pretrained-vectors.html.

[10] If a word occurs in both source and target, we use the word embedding of the source language.

	CROSS-LINGUAL		
Language	zero-shot	+fastText	+Morph
Galician	51.9	**72.8**	71.0
Kazakh	12.5	27.7	**28.4**
Kazakh (translit.)	21.2	31.1	**36.7**

Table 5.7: LAS results on the **development sets**. *zero-shot* denotes results where we predict using a model trained only on the source treebank.

			CROSS-LINGUAL		
	baseline	best system	+fastText	+Morph	rank
Galician	66.16	74.25	**70.46**	69.21	10/27
Kazakh (translit.)	24.21	31.93	25.28	**28.23**	2/27

Table 5.8: Comparison to CoNLL 2018 UD Shared Task on **test sets**. *best system* is the state-of-the-art model for each treebank: UDPipe-Future (Straka, 2018) for Galician and Uppsala (Smith et al., 2018) for Kazakh. *rank* shows our best model position in the shared task ranking for each treebank.

accuracy from 12.5 to 21.2 LAS. Cross-lingual training with MORPH further improves performance to 36.7 LAS.

5.5.2 *Comparison with CoNLL 2018*

To see how our best approach (i.e., cross-lingual model with MORPH augmentation) compares with the current state-of-the-art models, we compare it to the recent results from CoNLL 2018 shared task. Training state-of-the-art models may require lots of engineering and data resources. Our goal, however, is not to achieve the best performance, but rather to systematically investigate how far simple approaches can take us. We report performance of the following: (1) the shared task baseline model (UDPipe v1.2; Straka and Straková, 2017) and (2) the best system for each treebank, (3) our best approach, and (4) a cross-lingual model with fastText embeddings.

Table 5.8 presents the overall comparison on the test sets. For each treebank, we apply the same sentence segmentation and tokenization used by each best system.[11] We see that our approach outperforms the baseline models on both languages. For Kazakh, our model (with transliteration) achieves a competitive LAS (28.23), which would be the second position in the shared task ranking. As comparison, the best

[11] UD shared task only provides unsegmented (i.e., sentence-level and token-level) raw test data. However, participants were allowed to use predicted segmentation and tokenization provided by the baseline UDPipe model.

system for Kazakh (Smith et al., 2018) trained a multi-treebank model with four source treebanks, while we only use one source treebank. Their system use predicted POS as input, while ours depends solely on words and characters. The use of more treebanks and predicted POS is beyond the scope of our paper, but it is interesting that our approach can achieve the second best score with such minimal resources. For Galician, our best approach outperforms the baseline by 8.09 LAS points. Note that, the Galician treebank does not come with training data. We use 50:50 train/dev split, while other teams might use higher split for training (for example, the best system (Straka, 2018) uses 90:10 train/dev split). Since we treat Galician as a held-out data point, we did not tune on the proportion for training data, but we guess that this is the main reason why our system achieve rank 10 out of 27.

Compared to cross-lingual models with fastText embeddings (fastText vs. MORPH), we observe that our approach achieves better or comparable performance, showing its potential when there is not enough monolingual data available for training word embeddings.

5.6 CONCLUSIONS

In this paper, we investigated various low-resource parsing scenarios. We demonstrate that in the extremely low-resource setting, data augmentation improves parsing performance both in monolingual and cross-lingual settings. We also show that transfer learning is possible with *lexicalized* parsers. In addition, we show that transfer learning between two languages with different writing systems is possible, and future work should consider transliteration for other language pairs.

While we have not exhausted all the possible techniques (e.g., use of external resources (Rasooli and Collins, 2017; Rosa and Mareček, 2018), predicted POS (Ammar et al., 2016a), multiple source treebanks (Lim, Partanen, and Poibeau, 2018; Stymne et al., 2018), among others), we show that simple methods which leverage the linguistic annotations in the treebank can improve low-resource parsing. Future work might explore different augmentation methods, such as the use of *synthetic* source treebanks (Wang and Eisner, 2018) or contextualized language models (Devlin et al., 2018; Howard and Ruder, 2018; Peters et al., 2018) for scoring the augmented data (e.g., using perplexity).

Finally, while the techniques presented in this paper might be applicable to other low-resource languages, we want to also highlight the importance of understanding the characteristics of languages being studied. For example, we showed that although North Sámi and Finnish do not share vocabulary, cross-lingual training is still helpful because they share similar syntactic structures. Different language pairs might benefit from other types of similarity (e.g., morphological) and investigating this would be another interesting future work for low-resource dependency parsing.

ACKNOWLEDGMENTS

Clara Vania is supported by the Indonesian Endowment Fund for Education (LPDP), the Centre for Doctoral Training in Data Science, funded by the UK EPSRC (grant EP/L016427/1), and the University of Edinburgh. Anders Søgaard is supported by a Google Focused Research Award. We thank Aibek Makazhanov for helping with Kazakh transliteration, and Miryam de Lhoneux for parser implementation. We also thank Andreas Grivas, Maria Corkery, Ida Szubert, Gozde Gul Sahin, Sameer Bansal, Marco Damonte, Naomi Saphra, Nikolay Bogoychev, and anonymous reviewers for helpful discussion of this work and comments on previous drafts of the paper.

APPENDICES

A Effects of Fine-Tuning for Cross-Lingual Training

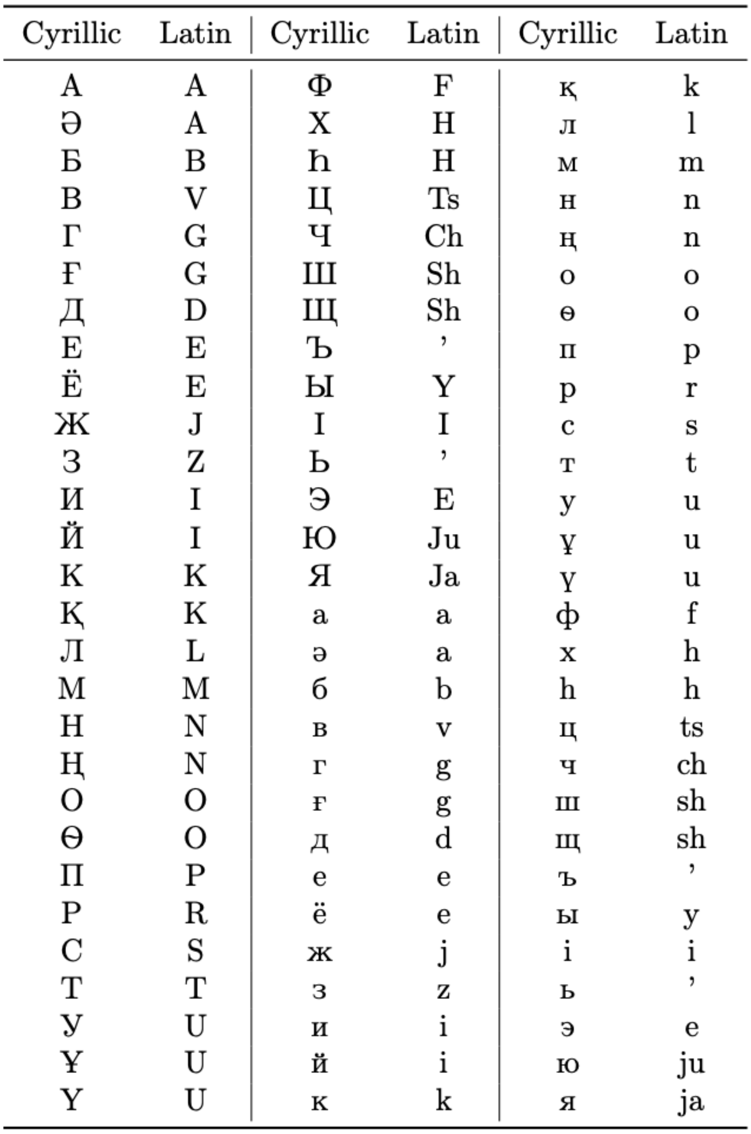
For our cross-lingual experiments in Section 5.2.3, we observe that fine-tuning on the target treebank always improves parsing performance. Table 5.9 reports LAS for cross-lingual models with and without fine-tuning.

size	cross-base	+Morph	+Nonce
$\mathcal{T}_{100}$	57.9 (+4.6)	59.5 (+6.2)	59.3 (+6.0)
$\mathcal{T}_{50}$	48.3 (+5.8)	49.8 (+7.3)	50.1 (+7.6)
$\mathcal{T}_{10}$	29.8 (+11.3)	34.9 (+16.4)	34.8 (+16.3)
	↓ *with fine tuning (FT)* ↓		
$\mathcal{T}_{100}$	61.3 (+8.0)	60.9 (+7.6)	**61.7 (+8.4)**
$\mathcal{T}_{50}$	52.0 (+9.5)	51.7 (+9.2)	**52.0 (+9.5)**
$\mathcal{T}_{10}$	34.7 (+16.2)	**37.3 (+18.8)**	35.4 (+16.9)

Table 5.9: Effects of fine-tuning on North Sámi development data, measured in LAS. *mono-base* and *cross-base* are models without data augmentation. % improvements over *mono-base* shown in parentheses.

B Cyrillic to Latin Alphabet mapping

We use the following character mapping for Cyrillic to Latin Kazakh treebank transliteration.



Cyrillic	Latin	Cyrillic	Latin	Cyrillic	Latin
А	A	Ф	F	қ	k
Ә	A	Х	H	л	l
Б	B	ћ	H	м	m
В	V	Ц	Ts	н	n
Г	G	Ч	Ch	ң	n
Ғ	G	Ш	Sh	о	o
Д	D	Щ	Sh	ө	o
Е	E	Ъ	’	п	p
Ё	E	Ы	Y	р	r
Ж	J	І	I	с	s
З	Z	Ь	’	т	t
И	I	Э	E	у	u
Й	I	Ю	Ju	ұ	u
К	K	Я	Ja	ү	u
Қ	K	а	a	ф	f
Л	L	ә	a	х	h
М	M	б	b	һ	h
Н	N	в	v	ц	ts
Ң	N	г	g	ч	ch
О	O	ғ	g	ш	sh
Ө	O	д	d	щ	sh
П	P	е	e	ъ	’
Р	R	ё	e	ы	y
С	S	ж	j	і	i
Т	T	з	z	ь	’
У	U	и	i	э	e
Ұ	U	й	i	ю	ju
Ү	U	к	k	я	ja

Figure 5.3: Cyrillic to Latin alphabet mapping.

6 THE APPOSCORPUS: A NEW MULTILINGUAL, MULTI-DOMAIN DATASET FOR FACTUAL APPOSITIVE GENERATION

ABSTRACT

News articles, image captions, product reviews and many other texts mention people and organizations whose name recognition could vary for different audiences. In such cases, background information about the named entities could be provided in the form of an appositive noun phrase, either written by a human or generated automatically. We expand on the previous work in appositive generation with a new, more realistic, end-to-end definition of the task, instantiated by a dataset that spans four languages (English, Spanish, German and Polish), two entity types (person and organization) and two domains (Wikipedia and News). We carry out an extensive analysis of the data and the task, pointing to the various modeling challenges it poses. The results we obtain with standard language generation methods show that the task is indeed non-trivial, and leaves plenty of room for improvement.

6.1 INTRODUCTION

News articles, image captions, product reviews and many other texts mention people and organizations, whose name recognition could vary for different audiences. A piece of news, for example, may concern people and organizations that are known locally, but are not necessarily well-recognized on a global level. In such cases, news pieces targeted at a wider audience would provide background information about the entity in focus, often in the form of an *appositive*. For example:

> In March 2017 , Natalie Jaresko, *former Minister of Finance in Ukraine*, was appointed as the board's executive director.

It is unlikely that many people outside of Ukraine know the name Natalie Jaresko, so a foreign reader would likely benefit from the extra bit of information about her former occupation as justification for her new appointment. An appositive could also be less contextualized and provide information of more general importance, for example:

> The conservation unit is in the Calhau bairro of São Luís, *the state capital*.

In general terms, appositives are phrases that appear next to a noun phrase and serve an explicative function (Bauer, 2017). Adding such explanations to text is a multi-step process. First, one has to decide whether an entity mention needs an appositive. That may not be the case for entities that are sufficiently well-known or that have been introduced earlier in the text. In case an appositive is indeed needed, the next step is to choose what information about the entity to disclose. If the information is to be of a factual nature, the writer needs to have prior knowledge of the entity, or access to an external knowledge resource–Kang et al. (2019) found appositives to be frequently based on facts of particular relevance to the context of the mention. Lastly, the surface form of the appositive, well-fitted to the surrounding context, needs to be produced. Viewed from the perspective of NLP, appositive generation is therefore an interesting and challenging natural language generation problem that involves reasoning over facts from an external knowledge source, with reference to a given context.

The task of appositive generation, first introduced by Kang et al. (2019), is still in its early stages and data resources are limited. We expand on previous work in appositive generation with a new, more realistic, end-to-end definition of the task, instantiated by a dataset,`ApposCorpus`,[1] that spans four languages (English, Spanish, German and Polish), two entity types (person and organization) and two domains (Wikipedia and News). While Wikipedia as a domain is curated for a world-wide audience and as such may not benefit much from appositive generation, we posit that it is a valuable source of abundant cross-lingual data which could be used as the basis for transfer learning. In addition to a large training set automatically sourced from Wikipedia, we therefore also introduce a gold standard test sourced from news wire, one of the true target domains for appositive generation (Kang et al., 2019).

6.2 THE TASK: APPOSITIVE GENERATION

Kang et al. (2019) laid the groundwork for appositive generation and our work can be seen as an expansion of their efforts. Yet, we both rename the task and redefine it in more general terms.

6.2.1 *Prior work*

Kang et al. (2019) introduced the task of appositive generation. To date this is the only work on this task. They designed a data collection procedure where appositives are identified by locating instances of the `appos` dependency label (Nivre et al., 2020) in parsed text, and used it to build a dataset of appositives for PERSON entities in English news articles. The candidate appositives were cross-referenced with

[1] Available at `https://yovakem.github.io/#ApposCorpus`.

the WikiData knowledge base (Vrandečić and Krötzsch, 2014) through word matching, and only those appositives were included in the final dataset which matched a fact from WikiData.

More generally, appositive generation relates to work on joint fact selection and generation (Angeli, Liang, and Klein, 2010; Kim and Mooney, 2010; Konstas and Lapata, 2013; Liang, Jordan, and Klein, 2009).

6.2.2 *A shift in terminology*

Kang et al. (2019) actually called the phrases in question *post-modifiers*, rather than appositives. The linguistic term *post-modifier* can be seen as subsuming appositives, but it is much broader, including also prepositional, non-finite and dependent clauses that appear in postposition. Meanwhile, appositives come in two forms, nominal appositives, where a single noun identifies or qualifies another noun, e.g. President *Washington*, and explicative appositives, where a pronoun, an infinitive or a noun phrase is used to explain or specify the status of a noun (Bauer, 2017). Explicative appositives are further characterized as *non-essential*, meaning that they are not integral to the grammatical or semantic well-formedness of the sentence it appears in, and as such are often delimited from the rest of the sentence by punctuation marks (Traffis, 2019). For the purposes of providing background information about named entities, we are in particular interested in *explicative appositive noun phrases*, and that is what we refer to as an *appositive* throughout this work.

6.2.3 *Expanding the task definition*

Being built with reference to WikiData, the dataset of Kang et al. (2019) creates the illusion that all facts necessary to generate an appositive are available in the knowledge base. Balaraman, Razniewski, and Nutt (2018) studied the relative completeness of WikiData entries and found gaps to be the norm rather than an exception. Moreover, the dataset of Kang et al. (2019) only includes positive samples, i.e. instances where an appositive is due. A more realistic scenario would also require the model to choose whether or not to add an appositive to a given entity mention. `ApposCorpus` is *not* constrained by WikiData in terms of fact matching, and contains positive *and* negative samples, i.e. instances of empty appositives. See Figure 6.1 for an illustration. Moreover, it is multilingual and covers both PERSON and ORGANIZATION named entities. We built this dataset primarily based on text from Wikipedia, chosen for its rich cross-lingual coverage.

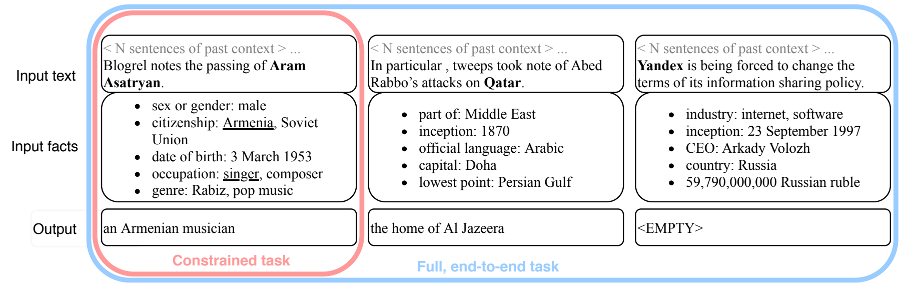

Input text
Input facts
Output
< N sentences of past context > ...
Blogrel notes the passing of **Aram Asatryan**.
sex or gender: male
citizenship: Armenia, Soviet Union
date of birth: 3 March 1953
occupation: singer, composer
genre: Rabiz, pop music
an Armenian musician
Constrained task
< N sentences of past context > ...
In particular , tweeps took note of Abed Rabbo's attacks on **Qatar**.
part of: Middle East
inception: 1870
official language: Arabic
capital: Doha
lowest point: Persian Gulf
the home of Al Jazeera
< N sentences of past context > ...
Yandex is being forced to change the terms of its information sharing policy.
industry: internet, software
inception: 23 September 1997
CEO: Arkady Volozh
country: Russia
59,790,000,000 Russian ruble
<EMPTY>
Full, end-to-end task

Figure 6.1: Illustration of the task in a *constrained* setting, where an appositive is always due and the facts in it are always available in the knowledge base; and in a *full, end-to-end* setting, where a decision has to be made as to whether or not to generate an appositive, and the facts in the appositive may be missing from the knowledge base. The entity in focus is shown in bold, the relevant facts are underlined (where available), and the <EMPTY> tag means no appositive is needed. Optionally, previous context can be included in the input, e.g. the three previous sentences–this is not shown in the figure.

6.3 DATASET COLLECTION: WIKIPEDIA

We used the March 2020 Wikipedia dump[2] for English, Spanish, German and Polish, which we parsed with WikiExtractor,[3] preserving internal links.[4] The choice of these particular languages was mostly based on the availability of good dependency parsers. Since dependency parsing is an integral step of the data collection process (Kang et al., 2019), it has to be as precise as possible, to maximize the quality of the outcome.[5] Below, we describe in detail the data collection procedure.

6.3.1 *Preprocessing*

We processed every article as follows: (1) tokenize the text and segment sentences; (2) normalize mentions of the entity in the article's title and annotate them with internal links; (3) identify sentences which contain a linked named entity listed as an instance of type *human* or of (a subclass of) type *organization* in WikiData (corresponding to the PERSON and ORGANIZATION named entity types); (4) run a dependency parser on these sentences. Steps (1) and (4) were performed with Stanza (Qi et al., 2020).

[2] https://dumps.wikimedia.org/
[3] https://github.com/attardi/wikiextractor
[4] These links point to other pages on Wikipedia and allow us to identify the Wikidata entry for the given named entity.
[5] We only considered parsers with labeled accuracy score over 90.0

6.3.2 *Detecting appositives*

Any instance of the appos label that depends on a linked named entity and is separated from it with a comma or an opening parenthesis was considered a valid candidate. In this case, we recorded the source sentence, replacing the appositive with special token <appos>, as input data, and the appositive as a target. The beginning of the appositive was taken to be the first token after the comma or opening parenthesis, and the end is taken to be the last token before the next comma/semicolon/full stop (if beginning was marked by a comma) or closing parenthesis (if beginning was marked by an opening parenthesis). We discarded any commas and parenthesis surrounding the appositive, but kept semicolons and full stops as part of the input sentence. We also recorded the three preceding sentences from the article and one following sentence. Similarly to Kang et al. (2019), we process one appositive per sentence, i.e. if there are multiple appositives in a sentence, we select the first one and do not consider the rest.

Appositives containing just dates (usually the date of birth and/or death of a person) are ubiquitous across Wikipedia articles to the point that they constitute up to 30% of the data samples that we get with the procedure described above. We reduced this imbalance in the data by downsampling this type of appositives to only 10% of its occurrences.

6.3.3 *Negative samples*

We added negative samples to the dataset, matching the number of positive ones. They were drawn according to the following criteria: (1) there is a PERSON or ORGANIZATION entity in the sentence, (2) it is not followed by a comma or opening parenthesis, and (3) the rest of the sentence does not contain an appositive dependent on the PERSON or ORGANIZATION entity. Condition (2) was used to reduce the chance of including instances of appositives that were not correctly tagged as such by the parser (recall that appositives are often delimited from the rest of the sentence by a punctuation), while (3) was used to ensure that we did not include instances that contain a non-essential appositive, which the author had failed to delimit by any punctuation. In negative samples the input is the original source sentence with an added token <appos> just after the PERSON/ORGANIZATION entity, and the target is a special <EMPTY> token.

The procedure described above was used to collect *training* data for factual appositive generation. As it is our goal to study the potential of a cross-domain approach to appositive generation, the ApposCorpus also contains an out-of-domain test set, sourced from news wire.

6.4 DATASET COLLECTION: NEWS

We sourced our data for cross-domain evaluation from the news domain, following previous work (Kang et al., 2019), using these news corpora: Global Voices (English, Spanish, German, Polish) (Tiedemann, 2012), News Commentary (English, Spanish, German) (Tiedemann, 2012) and Paralela (Pęzik, 2016).

6.4.1 *Entity linking*

Unlike Wikipedia, where entities are explicitly linked to WikiData entries through internal links, here, we had to perform additional entity linking. We did so in the following manner: (1) extract candidates from WikiData based on exact match between the full span of the named entity and all aliases of entities in the respective subset of WikiData (instances of type *human* if NER label is PERSON , else *organization*), (2) obtain relative alias frequency distributions from Wikipedia, and (3) the candidate entity with the highest relative frequency given the alias is selected. We chose to use this prior-based method instead of a modeling approach since off-the-shelf entity linkers were not available for all the languages involved. Candidate appositives were then identified as described in 3.2.

As errors could occur both in the entity linking and in the appositive detection, we hired manual annotators to verify the output of the two procedures (see details in Appendix A). The News portion of the `ApposCorpus` is therefore gold standard and will serve for stable, accurate evaluation.

6.4.2 *Negative samples*

We added $1,000 - n$ negative samples to each subset of the data, where n is the number of positive samples. The exact ratio of positive samples in each test set is reflected in the `always yes` baseline shown in the Results section (see Figure 6.2a), a dummy baseline which always predicts an appositive.

6.5 DATA ANALYSIS

This section outlines some findings on the properties of our dataset, based on general statistics and WikiData cross-referencing. The procedure described above yielded less than four thousand samples of ORG appositives for Polish, so this subset is omitted from the `ApposCorpus`.

		en	es	de	pl
PER	Size	559k	164k	269k	14k
	Length	3.4	4.16	2.7	2.67
	WD (%)	25.5	28.1	21.2	22.3
ORG	size	612k	104k	333k	-
	Length	2.19	4.09	1.64	-
	WD (%)	27.8	24.9	22.2	-

(a) Wikipedia

		en	es	de	pl
PER	size	1k	1k	1k	1k
	Length	4.07	3.51	3.43	2.32
	WD (%)	29.3	30.8	21.7	20.7
ORG	Size	1k	1k	1k	-
	Length	3.31	3.08	1.87	-
	WD (%)	35.4	30.0	22.8	-

(b) News

Table 6.1: Dataset statistics. *Size*: full dataset size, *Length*: average appositive length, *WD*: ratio of appositives matching a fact from WikiData.

6.5.1 *General statistics*

Table 6.1 lists some statistics about the two parts of the dataset, one based on Wikipedia (*Wikipedia data*) and the other on news (*News data*). Row *Size* refers to the full size of the data as split into language and entity type. We further split each Wikipedia subset for training (70%), validation (15%) and testing (15%). Size varies greatly across the data, with the Polish PERSON subset being merely 2.5% the size of the English PERSON subset. This relates both to a difference in the Wikipedia sizes for these languages (6M English articles v. 1.4M Polish articles) and to a difference in the frequency of use of appositives across the languages.

Row *Length* in Table 6.1 lists the average number of tokens per appositive, which varies from two to four tokens, and is generally lower for ORGANIZATION appositives than for PERSON ones.

6.5.2 *Cross-referencing with WikiData*

As discussed before, fact matching is not part of our data collection procedure, but at training time it would be beneficial to have access to a knowledge base such as WikiData, and to draw from it, when possible. So we extract the WikiData entries for all named entities in our dataset and perform word matching between facts and appositives in the following way: (1) tokenize the fact and the appositive, (2) remove stopwords, (3) measure token overlap. If the overlap is non-zero, we consider there to be a match and annotate the fact as `used`.[6]

Cross-referencing the data with WikiData is also useful as an insight into the makeup of the data, albeit an insight that is biased the scope and completeness of the knowledge base.

[6] We experimented with other thresholds (2 and 3-word overlap) and with fuzzy matching, but found this method to work best.

	Fact type	News (%)	Wiki(%)		Fact type	News (%)	Wiki(%)
per	position held	20.9	9.4	org	instance of	23.1	10.9
	occupation	15.9	10.6		official website	6.3	6.2
	citizenship	10.1	4.3		country	5.9	3.3
	member of party	7.6	1.9		member of	4.2	2.4
	award received	5.2	3.9		subsidiary	3.5	2.1
	nominated for	3.6	0.4		capital of	3.2	0.1
	educated at	3.1	3.1		has quality	3.0	0.0

Table 6.2: Top fact types. English.

coverage Row *WD* in Table 6.1a shows the rather low percentage of appositives from the Wikipedia dataset that are matched to at least one fact from WikiData: from 21.2 for German person appositives to 28.1 for Spanish person appositives. The numbers for the News test set, shown in Table 6.1b, are mostly similar to those for the WikiData. Another way to view these percentages is as an effective upper bound on the performance of a model trained with WikiData as the source of knowledge. Further work in identifying other sources of facts for appositive generation and new means of integrating them into a model could therefore prove very fruitful.

We manually inspected a random sample of 100 appositives from the English section of the Wikipedia dataset that were not matched to any fact from WikiData. In the majority of cases, the appositives concerned the occupation of a person, their position within an organization, their country of origin, or other type of information that is typically found in WikiData, but was missing for the given named entry.

composition We studied the composition of the data, as observed with reference to WikiData. We performed our analysis on all languages and found that similar trends hold cross-lingually, so here we discuss the English portion of the data only, and in the Appendix B we include the corresponding tables for Spanish, German and Polish.

We looked at the types of facts that were matched to appositives from the News data. For all fact types that constitute 3% or more of all facts matched, we also looked at their frequency in the Wikipedia data. Results are shown in Table 6.2. We see that half of the top fact types in the News dataset are also well-attested in the Wikipedia data, i.e. their relative frequency is 3% or more. We can expect that knowledge concerning appositives based on these fact types would trivially transfer from one domain to the other. The low frequency for the remaining fact types (cf. *has quality* and *capital of*), on the other hand, poses a challenge whose solution would require deeper natural language understanding and, possibly, explicit domain transfer techniques.

6.6 experiments

To show how the new task formulation can be used, we experiment with three established language generation methods: the main method of Kang et al. (2019), which we refer to as `base`; an extension of `base` with external knowledge injected through embeddings with knowledge-base grounding (`KB`); and a model enhanced with an explicit copy mechanism (`copynet`, Gu et al. (2016)). Notice that our goal here is not to build the best model for this task, but to develop reasonable models which can serve as baselines for future work in this area.[7]

6.6.1 *Architectures*

lstm baseline, `base` Kang et al. (2019) introduced an LSTM-based encoder-decoder architecture with an auxiliary objective used to guide the attention of the decoder towards the WikiData facts that were matched during the data collection process. Input sentences and facts are represented with the same word embeddings and encoded by separate biLSTMs. The decoder is initialized with the encoding of the input and attends over the encodings of the facts. Our only modification here is to add a "None of the above" item to the list of facts about an entity and point the attention to that when no other fact was matched or the appositive was empty (i.e. for negative data instances).

lstm with external knowledge, `kb` External knowledge can be beneficial to a better understanding of the context and how it relates to the different facts known about an entity. Here, we use the same architecture as above, but initialize the embedding matrix of the model with the NTEE (Neural Text-Entity Encoder) word embeddings, trained on Wikipedia with WikiData grounding (Yamada et al., 2017). They aim to represent a text and its relevant entities close to each other. We deem these embeddings suitable for our modeling setup, where input text and facts are represented in a shared space. Unfortunately, the NTEE embeddings are available only for English. So we used word-level translation to "project" them to Spanish, German and Polish. See more details in Appendix D. This approach is likely to introduce some noise, but we only use the projection to initialize the embedding matrix which is then further trained. So any signal coming from the embeddings can be used by the model and any noise can be filtered out during training.

[7] We also experimented with a transformer architecture, but encountered optimization problems. See details in Appendix E.

lstm with a copy mechanism, `copynet` Motivated by the observation that there is an overlap of at least one token between WikiData facts and appositives for about 25% of the datapoints in our dataset, we experiment with a method that allows the decoder to copy tokens directly from the input: Copynet (Gu et al., 2016). Kang et al. (2019) correctly point out that in their constrained data setting, where data points were selected based on word overlap with WikiData facts, using a copy mechanism would result in double-counting, i.e. artificially boosted results. In our data setting, however, this is not the case.

All three approaches are end-to-end in the sense that we do not split up the classification task of whether or not to predict an appositive from the task of generating an appositive where it is due. As negative samples in the dataset have the special `<EMPTY>` token as target, the models are performing the classification task *implicitly* by choosing whether to predict the `<EMPTY>` token or not.

Preliminary experiments with the `base` architecture showed that the choice between providing the model with three sentences of preceding context, one or zero had little impact on its performance, so all results reported below use one sentence of preceding context, following Kang et al. (2019). Further details on the implementations and the hyperparameters we used can be found in Appendix C.

6.6.2 *Evaluation*

We follow Kang et al. (2019) in the choice of performance metrics for predictions over the positive instances in the data: we measure F1 score of the predicted bag-of-words excluding stopwords; BLEU (Papineni et al., 2002) over n-grams, where $n = 1, 2, 3$;[8] and METEOR (Denkowski and Lavie, 2014), which supports stemming and synonymy only in English, Spanish and German, so these features are not used for Polish. We use accuracy to measure the models' ability to determine when an appositive is due.

6.7 results

We view the results of our experiments from two angles: one concerns the expansion of the task definition we achieve with `ApposCorpus`, from a constrained scenario to an end-to-end one; the other concerns the increased coverage of the dataset, which allows us to compare and contrast appositive generation across different languages and named entity types.

[8] Kang et al. (2019) also included four-grams, but seeing that the average length of an appositive across the different subsets of the data is 3.08 tokens, we exclude four-grams from consideration.

Train setting	Test setting	Dataset	Acc(%)	F1	BLEU	MET.
constrained	constrained	ApposCorpus (News)	-	19.61	7.93	9.12
		PoMo	-	11.21	3.44	5.03
end-to-end	constrained	ApposCorpus (News)	95.0	10.76	3.39	4.41
		PoMo	91.7	4.52	0.57	2.03
	end-to-end	ApposCorpus (News)	72.33	5.97	1.03	2.96

Table 6.3: Generation of English PERSON appositives in a constrained v. end-to-end train and test setting.

6.7.1 *Constrained v. end-to-end scenario*

To draw a direct comparison to the work of Kang et al. (2019), in this subsection we focus on English PERSON appositives, as this is the subset that was covered by their dataset, dubbed `PoMo`. We begin by replicating exactly their train and test settings, both constrained, using the model architecture they proposed, `base`. In the first two rows of Table 6.3, the performance of the model is reported on both the constrained subset of our News test data and on the `PoMo` test set, constrained by design. There is a considerable difference in performance as measured on the two test sets. Since they were both drawn from the same domain, this difference may largely be due to one test set being gold standard and the other silver standard, which highlights the importance of having gold standard evaluation data.

Using these results as a starting point, we consider two important factors in the shift from a constrained to an end-to-end setting: one concerns learning and the other, evaluation.

LEARNING COMPLEXITY can be expected to increase in the end-to-end training setting, since the model has to learn not just what appositive to predict, but also whether or not to predict an appositive. Due to gaps in WikiData, the model also has to learn how to best handle instances of appositives based on unobserved facts. We demonstrate how these factors affect performance by comparing the `base` model trained in a constrained setting to one trained in an end-to-end setting. We measure the models' performance in a constrained test setting, to make the comparison fair to the former. Shifting from a constrained train setting (rows 1 and 2 of Table 6.3) to an end-to-end setting (rows 3 and 4), we observe a drop in performance of around 50% on all generation metrics. The model trained end-to-end does very well on choosing whether or not to predict an appositive (accuracy is 95% for `ApposCorpus` and 91.7% for `PoMo`) , so we have to conclude that the lower generation scores are not a matter of predicting empty appositives, but rather of predicting worse appositives due to the increased learning complexity.

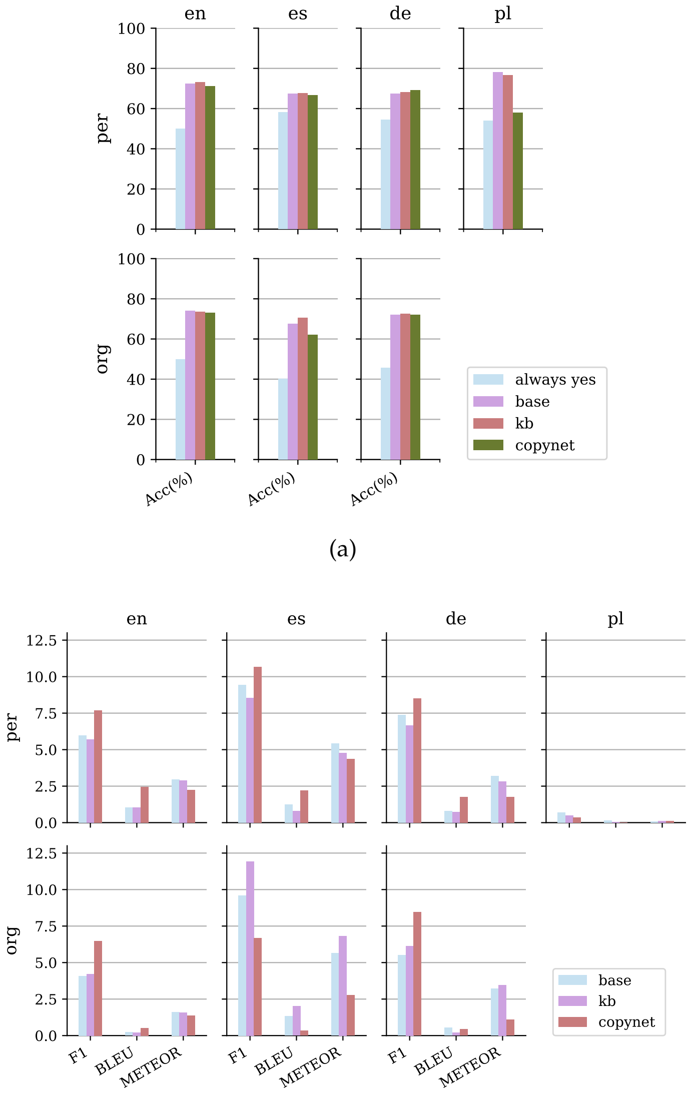
EVALUATION is another aspect to consider when comparing the constrained and full data settings. The quality of evaluation is key to understanding how well a model would perform if deployed in the real world. Constrained evaluation, however, only tells us how a model would do in an idealistic scenario, where all the facts about all the entities were indeed covered by a knowledge base. As this is not the case with WikiData (Balaraman, Razniewski, and Nutt, 2018), and with any existing knowledge base for that matter, it is important to evaluate models in a manner that reflect gaps in external knowledge sources. We report the performance of a model trained and tested in an end-to-end setting in the last row of Table 6.3. Compared to the model's performance as measured on the constrained test set (row 3), these numbers are substantially lower. Yet, they are the numbers that most truly represent the performance of the base model, at least in terms of automatic evaluation. We return to this matter when analysing the model's performance in Section 6.8.

6.7.2 *Languages and entity types*

The full range of results on the News test set are shown in Figure 6.2. Results are averaged over three models trained from different random initializations. We evaluate how well the models can detect when an appositive is needed (implicit classification) and how well it can perform the end-to-end task of classifying and generating a good appositive.

IMPLICIT CLASSIFICATION of the positive and negative samples in the data appears to be roughly equally challenging across the two entity types and the four languages, as viewed across all three models (see Figure 6.2a. One exception is Polish PERSON appositives, where `base` and `kb` score higher than they do on the other subsets, while `copynet` barely beats the baseline. Since the models are not explicitly trained to perform this type of classification, it is encouraging to see that even in this setting, they can outperform the baseline (always predicting the positive class) by as much as 20% in several cases.

GENERATION is the more difficult aspect of the task, as shown in Figure 6.2b. We see that, somewhat surprisingly considering the amounts of training data (see Table 6.1a), performance is not highest on English, but rather on Spanish. In line with the small amount of training data, on the other hand, performance on Polish is virtually non-existent. There are no clear differences between overall performance on PERSON and ORGANIZATION appositives. Only in English, the latter seem to pose a greater challenge to all three models and according to all three metrics. Model comparison is not straightforward since the different metrics reveal different strengths and weaknesses



(a)

(b)

Figure 6.2: Evaluation of (a) the models' ability to correctly decide when an appositive is due, (b) generated predictions for positive test instances. Measured on the News test set.

	true	system	neutral		true	system	neutral
PER	46.3%	47.3%	6.1%	ORG	66.0%	26.5%	7.4%

Table 6.4: Results from the ranking paradigm study comparing true appositives to system-generated ones.

in each approach. It does appear to be the case that injecting external knowledge through pretrained knowledge-base embeddings (kb) is beneficial to the prediction of ORGANIZATION appositives and somewhat harmful to the prediction of PERSON appositives. Since the differences between the three methods are not consistent across all languages, named entity types and metrics, we cannot conclusively say which method is best, but we do note that copynet scores high on the most metrics, languages and entity types. To better understand the performance of this model, we stepped away from automatic generation metrics, which are known to suffer from certain biases and can be difficult to interpret, and we carried out an additional manual evaluation.

6.8 ANALYSIS

6.8.1 *Manual evaluation*

We used Amazon Mechanical Turk to carry out a ranking paradigm study on the predictions of copynet for English PERSON and ORGANIZATION appositives. The annotators were shown a true appositive and a predicted appositive side-by-side (in the context of the input sentence) and were asked to express their preference towards either of these two, or their lack of preference as a third option. One example prompt is shown in Appendix F. Five annotations were obtained per data instance, and we then took the majority vote as an indication of the overall preference. The results are shown in Table 6.4. For PERSON appositives, the writer's choice (true) and the system's prediction were preferred at almost equal rates—this observation challenges the numbers obtained with the automatic evaluation metrics, as it suggests that the predicted appositives were not necessarily of poor quality. A bigger gap was observed between true and system-generated ORGANIZATION appositives, where the crowdworkers preferred the original appositive 66.0% of the time. The lower preference for ORGANIZATION predictions is in line with the trend in the automatic results, where performance on English ORGANIZATION appositives was shown to be lower than that on English PERSON appositives. Notice, however, that even for organization appositives, annotators still showed preference for the predicted ones at a considerable rate. This suggests that the automatic metrics may have severely under-represented the abilities of the models.

	Gold sentence	Prediction
1	In response to an April 9 court ruling declaring the military backed government of Frank Bainimarama <EMPTY> to power illegally when he dissolved Parliament and deposed the government of Laisenia Qarase , the country 's President nullified the Fiji 's constitution , fired the entire judiciary and appointment himself head of state and the armed forces.	the President of Fiji
2	Artyom Loskutov, creator of the popular counter - culture art movement " Monstration ", made waves on RuNet by signing a letter in support of Dmitry Kiselyov , a journalist who many consider to be Putin 's chief propagandist.	a Russian painter
3	In a fatal blow to our already lackluster sources of entertainment , the Sudanese government has blocked access to YouTube, the online video sharing Web site.	platform
3	Modi, a tech-savvy nationalist from the right-wing Bharata Janatiya Party, has traveled the world to sell the idea of India as an emerging digital economy, making deals with the likes of Google and (less successfully) Facebook.	the Prime minister of India
4	Some critics also highlighted the fact that Jabrailov is from Chechnya, a republic in the Northern Caucasus region of Russia where Muslim separatists fought two bloody wars against the Russian army.	Chechnya
5	He steers between Soukous , rhumba and RnB " , and links to an interview with the singer on Radio Okapi, the nationwide radio station sponsored by the UN and Fondation Hirondelle.	the Democratic Republic of the Congo
6	In particular , tweeps took note of Abed Rabbo 's attacks on Qatar, the home of Al Jazeera.	Qatar

Table 6.5: Examples where true appositives were preferred over predicted ones by human annotators.

6.8.2 *Qualitative analysis*

To better understand the source of error in the models' predictions, we manually inspected 50 data points from the PERSON subset and 50 data points from the ORGANIZATION subset, where a choice was made in favour of the true appositive over the predicted one. Half of the data points were instances where the true appositive was not empty, but the models predicted an empty appositive. The annotators seemed to strongly prefer non-empty appositives, possibly due to the fact that they where shown sentences without their original context, where the role of an entity might have been clarified at an earlier mention. Yet, that is not categorically so as seen in example 1 in Table 6.5, where the predicted appositive is redundant in the given context, so the annotators preferred the true empty appositive. Other types of errors the models made were to predict appositives that concern the right piece of information but are too general (examples 2 and 3), to predict appositives based on the wrong piece of information (examples 4 and 5), and, specifically for ORGANIZATION appositives, to just repeat the named entity (example 6). While the latter is the result of either suboptimal learning or noise in the data, the rest of the errors we saw point to the need for an approach with deeper understanding of the facts and their relevance to the context.

6.9 conclusion

ApposCorpus targets factual appositive generation, a phenomenon frequently occurring in a range of textual domains. It substantially extends the prior resources in the area by spanning four languages, two named entity types , and two domains. This resource also allows the burgeoning field to investigate end-to-end appositive generation. Our manual and automatic evaluations with ApposCorpus show that standard model architectures can approach the quality of human targets in specific cases but there is still room for improvement. With this dataset, appositive generation can be studied in much more depth than previously possible, ultimately paving the way for novel NLP applications in the generation and writing space. The focus in future research, we believe, should fall on explicit methods for cross-domain learning, on richer knowledge sources, and on the development of test sets for new domains.

acknowledgements

We acknowledge the responsiveness and help of Jun Seok Kang and Robert L. Logan IV in discussions on the design choices and goals of their work on post-modifier generation and on how to accurately reproduce their experiments. We also thank researchers from Dataminr: Saran Krishnasamy, Lin Nie, and Isabel Zhang for their help in setting up experiments and annotation. Finally, we thank the three anonymous reviewers for their feedback and comments.

appendices

A Manual validation

We hired annotators fluent in the languages of the data to manually validate it. They had to mark instances where an error had occurred in the appositive detection, the detected appositive was not factual, or the entity had been incorrectly linked to WikiData (all instances of *noise* in the data, resulting from errors in appositive detection and entity linking). Our ultimate goal was to build test sets of 1,000 instances per language per entity type, equally balanced between positive and negative instances, i.e. we needed 500 valid data instances per language per entity type. Based on a pilot study, we determined that noise levels for candidate appositives for person entities were approximately 33% and for organization entities, 50% (averaged across languages). We therefore gave annotators 750 and 1,000 candidates to annotate for the person and organization types, respectively. For most language-entity type combinations, the manual annotation successfully yielded close to 500 valid instances. That was not the case for Polish orga-

	Fact type	News (%)	Wiki(%)
PER	position held	20.9	9.4
	occupation	15.9	10.6
	citizenship	10.1	4.3
	member of party	7.6	1.9
	award received	5.2	3.9
	nominated for	3.6	0.4
	educated at	3.1	3.1
ORG	instance of	23.1	10.9
	official website	6.3	6.2
	country	5.9	3.3
	member of	4.2	2.4
	subsidiary	3.5	2.1
	capital of	3.2	0.1
	has quality	3.0	0.0

Table 6.6: Top fact types in the English data.

nization appositives, where only 80 valid candidates were retrieved, so we excluded this language-entity type combination from our work. It remains an open question whether organization appositives in Polish are rare or our automatic detection method failed at catching them.

B Composition of cross-lingual data

Tables 6.6-6.9 present findings on the composition of the Spanish, German and Polish positions of the data, respectively, as observed through cross-referencing with WikiData. The low number of facts for Polish is the result of one fact type dominating a large amount of the data (*position held*).

C Implementations and hyperparameters

[base] and [kb] We use the implementation of Kang et al. (2019) from `https://github.com/rloganiv/claimrank-allennlp`. We set the model hyperparameters to the ones reported in their paper, changing only the dimension of the embeddings from 500 to 300, to make the comparison between [Base] and [KB] fair in terms of model parameterization. Training hyperparameters were tweaked to achieve stable training that fits on one 16 GB GPU. See the full list of hyperparameters in Table 6.10.

	Fact type	News (%)	Wiki(%)
PER	position held	27.1	4.9
	occupation	11.2	10.3
	citizenship	9.8	4.8
	participant of	8.7	1.1
	member of party	7.3	1.3
	award received	4.1	3.2
	employer	3.2	0.5
ORG	instance of	28.9	10.5
	country	7.8	3.5
	has quality	5.1	0.1
	capital of	4.4	0.5
	member of	4.3	2.9
	is located in	3.4	2.2

Table 6.7: Top fact types in the Spanish data.

	Fact type	News (%)	Wiki(%)
PER	position held	35.2	6.7
	employer	5.9	2.9
	citizenship	5.9	1.3
	member of party	5.3	2.4
	award received	4.9	2.1
	occupation	4.8	3.6
	participant of	4.6	1.5
	member of	3.6	1.0
ORG	official website	15.6	12.4
	instance of	15.1	4.5
	owner of	7.5	2.4
	member of	5.8	0.9
	has quality	4.6	0.0
	Commons category	3.2	4.5

Table 6.8: Top fact types in the German data.

	Fact type	News (%)	Wiki(%)
PER	position held	77.2	5.7
	participant of	3.2	0.7

Table 6.9: Top fact types in the Polish data.

	Base/KB	Copynet
vocab size	50k	50k
embedding dim	300	300
hidden units	250	250
num layers	2	2
optimizer	Adam	Adam
learning rate	0.001	0.0001
batch size	16	6
dropout	0.3	0.3

Table 6.10: Hyperparameter configurations for Base/KB models and Copynet models.

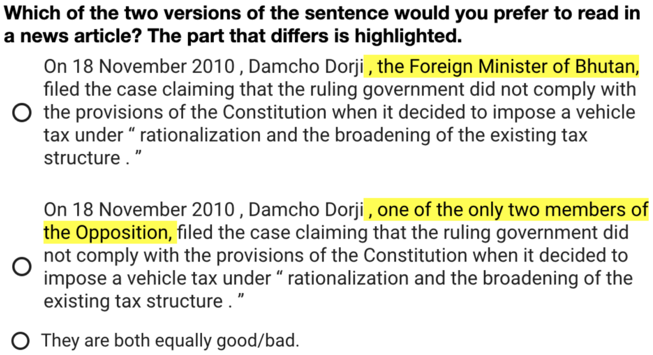
D Projection of NTEE embeddings

We obtained a bilingual dictionary with CSLS retrieval over the cross-lingual FastText embeddings (Bojanowski et al., 2017). CSLS retrieval is similar to nearest neighbor retrieval, but has proven more accurate: (Joulin et al., 2018b) report an accuracy of 83.7%, 77.6% and 73.5% for word translation, respectively, from Spanish, German and Polish to English, as measured on a sample of 1,500 medium frequency words. Any errors in the bilingual dictionaries would inevitably lead to noise in the NTEE embedding projection.

COPYNET We use the AllenNLP (Gardner et al., 2018) implementation of Copynet with the hyperpameters shown in Table 6.10.

E Transformer experiments

Transformer-based architectures are state-of-the-art for many NLP tasks, so it is only fair that we experiment with such an architecture as well. As BERT models (Devlin et al., 2019) have been made available for all four languages we work with, we chose to train BERT-to-BERT encoder-decoder models for appositive generation. Rothe, Narayan, and Severyn (2020) found that architecture to give strong performance on tasks like sentence fusion and rephrasing. We used their training schedule but unfortunately, found that all models learned to predict the `<empty>` token exclusively. As it is not the goal of our work to explore the capabilities of the BERT-to-BERT architecture in particular, we did not use further resources to adjust the training schedule. Yet, we do believe this to be an optimization problem, and we would not discourage future research from attempting to solve the task of appositive generation with a transformer-based approach.



Which of the two versions of the sentence would you prefer to read in a news article? The part that differs is highlighted.

○ On 18 November 2010 , Damcho Dorji , the Foreign Minister of Bhutan, filed the case claiming that the ruling government did not comply with the provisions of the Constitution when it decided to impose a vehicle tax under " rationalization and the broadening of the existing tax structure . "

○ On 18 November 2010 , Damcho Dorji , one of the only two members of the Opposition, filed the case claiming that the ruling government did not comply with the provisions of the Constitution when it decided to impose a vehicle tax under " rationalization and the broadening of the existing tax structure . "

○ They are both equally good/bad.

Figure 6.3: Prompt for manual evaluation.

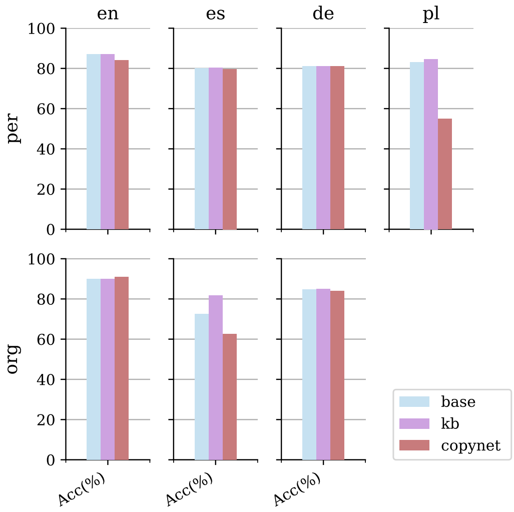
F Ranking paradigm study

Figure 6.3 shows an example prompt from the blind taste test. Instances where either the true appositive was empty or the predicted one was empty were included in the the study, but instances where both were empty were excluded, as the comparison would not have been meaningful in this case. The average time for completing a HIT was 53 seconds.

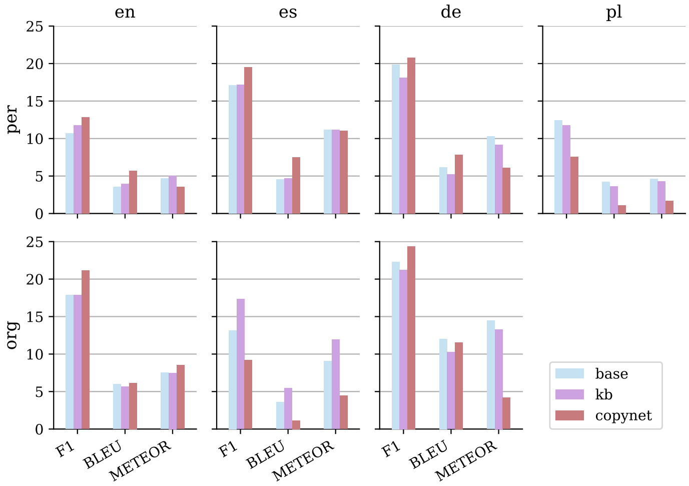
G Results

The results as measured on the Wikipedia test set are shown in Figure 6.4. Compared to results on the News test set (see Figure 6.2, the numbers seen here are higher, which is to be expected considering that this test set is in-domain and any noise found in it (due to it being silver standard) likely resembles the noise found in the training data. It is worth noting though, that certain patterns repeat between the two test sets, as for example the fact that `copynet`, as measured on F1 score and BLEU, outperforms the other models on the majority of language-named entity type combinations, but not on Polish PERSON appositives and Spanish ORGANIZATION appositives. This suggests that, while evaluation on the silver-standard Wikipedia test set cannot be consider fully stable and representative, it can be taken as a proxy in model comparison for developmental purposes.

The numbers behind the results from Figures 6.2 and 6.4 are shown in Tables 6.11 and 6.12, respectively.



(a)



(b)

Figure 6.4: (a) Evaluation of (a) the models' ability to correctly decide when an appositive is due, (b) generated predictions for positive test instances. Measured on the News test set.

		Acc	F1	BLEU	METEOR
EN-PER	always yes	0.5	0.0	0.0	0.0
	base	71.91	0.72	1.03	2.96
	kb	71.73	0.73	1.03	2.9
	copynet	70.19	0.71	2.45	2.24
EN-ORG	always yes	0.5	0.0	0.0	0.0
	base	68.09	0.74	0.23	1.61
	kb	66.58	0.74	0.21	1.58
	copynet	65.98	0.73	0.52	1.37
ES-PER	always yes	0.58	0.0	0.0	0.0
	base	62.84	0.67	1.24	5.44
	kb	61.18	0.68	0.79	4.79
	copynet	64.08	0.67	2.2	4.36
ES-ORG	always yes	0.4	0.0	0.0	0.0
	base	47.39	0.68	1.32	5.64
	kb	60.91	0.71	2.02	6.83
	copynet	33.57	0.62	0.33	2.77
DE-PER	always yes	0.54	0.0	0.0	0.0
	base	62.94	0.67	0.8	3.19
	kb	63.47	0.68	0.75	2.84
	copynet	67.63	0.69	1.75	1.76
DE-ORG	always yes	0.46	0.0	0.0	0.0
	base	62.77	0.72	0.54	3.21
	kb	65.13	0.73	0.19	3.47
	copynet	62.22	0.72	0.43	1.08
PL-PER	always yes	0.54	0.0	0.0	0.0
	base	72.83	0.78	0.16	0.09
	kb	73.8	0.77	0.06	0.13
	copynet	2.75	0.58	0.0	0.12

Table 6.11: Evaluation of the models' ability to correctly decide when an appositive is due and of generated predictions for positive test instances. Measured on the News test set.

		Acc	F1	BLEU	METEOR
EN-PER	base	85.42	0.87	3.59	4.7
	kb	85.32	0.87	3.94	5.07
	copynet	81.95	0.84	5.7	3.57
EN-ORG	base	89.47	0.9	5.99	7.54
	kb	89.43	0.9	5.65	7.49
	copynet	89.17	0.91	6.17	8.53
ES-PER	base	78.49	0.8	4.57	11.17
	kb	78.73	0.8	4.71	11.19
	copynet	79.1	0.8	7.54	11.04
ES-ORG	base	61.63	0.73	3.58	9.07
	kb	80.69	0.82	5.45	11.92
	copynet	46.91	0.62	1.14	4.46
DE-PER	base	79.28	0.81	6.19	10.35
	kb	80.09	0.81	5.24	9.19
	copynet	80.12	0.81	7.85	6.11
DE-ORG	base	84.09	0.85	12.04	14.45
	kb	84.58	0.85	10.28	13.31
	copynet	82.07	0.84	11.57	4.19
PL-PER	base	82.67	0.83	4.23	4.61
	kb	84.46	0.85	3.67	4.29
	copynet	24.49	0.55	1.1	1.71

Table 6.12: Evaluation of the models' ability to correctly decide when an appositive is due and of generated predictions for positive test instances. Measured on the Wiki test set.

7 PUZZLING MACHINES: A CHALLENGE ON LEARNING FROM SMALL DATA

ABSTRACT

Deep neural models have repeatedly proved excellent at memorizing surface patterns from large datasets for various ML and NLP benchmarks. They struggle to achieve human-like thinking, however, because they lack the skill of iterative reasoning upon knowledge. To expose this problem in a new light, we introduce a challenge on learning from small data, *PuzzLing Machines*, which consists of *Rosetta Stone* puzzles from Linguistic Olympiads for high school students. These puzzles are carefully designed to contain only the *minimal* amount of parallel text necessary to deduce the form of unseen expressions. Solving them does not require external information (e.g., knowledge bases, visual signals) or linguistic expertise, but meta-linguistic awareness and deductive skills. Our challenge contains around 100 puzzles covering a wide range of linguistic phenomena from 81 languages. We show that both simple statistical algorithms and state-of-the-art deep neural models perform inadequately on this challenge, as expected. We hope that this benchmark, available at `https://ukplab.github.io/PuzzLing-Machines/`, inspires further efforts towards a new paradigm in NLP—one that is grounded in human-like reasoning and understanding.

7.1 INTRODUCTION

It is beyond doubt that recent deep learning (DL) models have become a real success story for the ML and NLP communities. Previous work has shown that these DL models are good at tasks which humans consider fast and intuitive/automatic, such as object detection and document classification. Such tasks require training on large labeled datasets and are handled by the so called `System1`, following the commonly used terminology by Kahneman (2011). However, human-level understanding involves another mechanism: the slow, rational and sequential `System2` (Kahneman, 2011), that enables learning with fewer samples—achieved via *reasoning with the right abstractions*. Recent debates and research on the shortcomings of deep learning models (Bengio, 2020; LeCun, 2020; Marcus, 2020; McClelland et al., 2019) have emphasized the importance of `System2` skills and reached a consensus on expanding DL models with `System2` abilities being one of the next big challenges. To foster research in this promising

Chikasaw	**English**
1. Ofi'at kowi'ã lhiyohli.	The dog chases the cat.
2. Kowi'at ofi'ã lhiyohli.	The cat chases the dog.
3. Ofi'at shoha.	The dog stinks.
4. Ihooat hattakã hollo.	The woman loves the man.
5. Lhiyohlili.	I chase her/him.
6. Salhiyohli.	She/he chases me.
7. Hilha.	She/he dances.
Now you can translate the following into Chickasaw:	
	The man loves the woman.
	The cat stinks.
	I love her/him.
Translate the following into English:	
Ihooat sahollo.	
Ofi'at hilha.	
Kowi'ã lhiyohlili.	

Table 7.1: The "Chickasaw" puzzle (Payne, 2005)

direction, we propose a unique challenge on learning from small data named *PuzzLing Machines* based on the Linguistic Olympiads—one of the 13 recognized International Science Olympiads targeted at high-school students. Kahneman (2011) discusses the two modes of human thinking which perfectly encapsulate the current (so called System1) and the desired state (System1+System2) of the deep learning field. System1 handles tasks that humans consider fast, intuitive and automatic, such as object detection and document classification. Recent deep learning (DL) models have shown great promise at this type of tasks—thanks to large training datasets. Yet, it is through slow, rational and sequential mechanisms that human-like abstract reasoning happens, to enable learning from just a few examples. This System2-style modeling is still in its early stages in DL, but is recognized as a much needed next step in the field (Bengio, 2020; LeCun, 2020; Marcus, 2020; McClelland et al., 2019). To foster research in this promising direction, we propose a unique challenge on "learning from small data": *PuzzLing Machines*, based on the Linguistic Olympiads—one of the 13 recognized International Science Olympiads targeted at high-school students.

The *PuzzLing Machines* challenge is based on one of the most common puzzle types in the Linguistic Olympiads: the *Rosetta Stone* puzzles (Bozhanov and Derzhanski, 2013), a.k.a. translation puzzles. An example is given in Table 7.1.[1] Although these puzzles take the form of a traditional "machine translation" task, they are different in many ways: Rosetta Stone puzzles contain a minimal, carefully

[1] Copyright University of Oregon, Department of Linguistics.

designed set of parallel expressions (words, phrases or sentences) in a foreign and in a familiar language (e.g., Chickasaw-English). This minimal set is *just* enough to deduce the underlying translation model, which typically involves deriving mini-grammar rules, extracting a lexicon, and discovering morphological and phonological rules. The actual task then is to translate new expressions—generally in both directions—using the model deduced from the parallel data. The assignments are carefully designed so that the expressions cannot be generated through simple analogy, but rather through the application of the discovered rules. These properties distinguish the *PuzzLing Machines* challenge from the modern MT task, as it relies on deductive reasoning with linguistic concepts that are central to System2, rather than exploiting statistical properties from large datasets as in System1.

The lack of reasoning skills of statistical systems has recently gained a lot of attention. Various datasets that require a wide range of background knowledge and different types of reasoning abilities have been introduced, such as ARC (Clark et al., 2018), GQA (Hudson and Manning, 2019a), GLUE benchmarks (Wang et al., 2018) and SWAG (Zellers et al., 2018). Our challenge distinguishes from previous benchmarks with some key properties. First, most of these reasoning tasks require external scientific or visual knowledge, which makes it hard to measure the actual reasoning performance. On the other hand, our challenge does not rely on any external, multimodal or expert-level information. Second, and more importantly, *PuzzLing* challenge consists of a minimal set of examples required for solution. That means, there exists no extra training data, ensuring that exploiting surface patterns would not be possible unlike in some of existing benchmarks (Gururangan et al., 2018).

In summary, this paper introduces a unique challenge, *PuzzLing Machines*, made up of ∼100 Rosetta Stone, a.k.a translation puzzles covering 81 languages from 39 different language families based on the Linguistic Olympiads. The challenge requires System2 skills—sequential reasoning and abstraction of linguistic concepts, discussed in detail in §7.2. We discuss the dataset and the linguistic phenomena in the resulting dataset supported with statistics and examples in §7.3. In §7.4, we present the results of intuitive baseline methods and strong MT baselines such as Transformers encoder-decoder (Vaswani et al., 2017) with integrated pretrained language models as applied to these puzzles. We show that, unsurprisingly, the puzzles cannot be easily or robustly solved by currently existing methods. We hope that this benchmark is going to evoke development of new deep MT/NLP models that operate in a human-like manner and reason upon linguistic knowledge, providing a new future research direction for NLP.

7.2 META-LINGUISTICS

Meta-linguistics is defined by Chomsky (1976) as "the knowledge of the characteristics and structures of language" as realised on the level of phonology, morphology, syntax and semantics. Any English speaker would likely have the linguistic capacity to produce the word *undo* when asked "What is the opposite of *do*?" Only a speaker with some level of meta-linguistic awareness, however, would further be able to reflect on the structure of the word they have produced: to identify *un-* as a unit that serves to negate words, to spot its similarity in function to other units like *dis-* and *de-*. He/she would also be aware that *un-* is not interchangeable with *dis-* and *de-*, since it attaches to the front of verbs and adjectives but not to nouns.

Meta-linguistic awareness is especially useful (and often improved) in the process of learning a new language, as it allows the learner to compare and contrast the structure and characteristics of the new language to those that he/she is already familiar with. It is desirable that systems for natural language processing possess meta-linguistic awareness, too, as that could hugely improve their cross-lingual generalizability, a problem that remains open after being approached from various engineering perspectives, often with little recourse to linguistics. However, measuring the meta-linguistic awareness of a system is not trivial. Existing probing techniques are mostly designed to measure how well neural models capture specific linguistic phenomena, e.g., whether a specific layer of an English language model can capture that *undo* is negative, instead of testing for meta-linguistic awareness. Our challenge takes a step further and tests whether the model can apply the underlying morphological processes, e.g. of verbal negation through prefixing. In addition, our challenge spans a wide-range of language families and covers a variety of linguistic phenomena (see §7.3.1), that qualifies it as a favorable testbed for measuring meta-linguistic awareness.

Let us demonstrate how meta-linguistic reasoning skills are used to solve the "Chickasaw puzzle" given in Table 7.1. The translation model is iteratively deduced as follows: (1) the word order in Chickasaw is Subject-Object-Verb (SOV), unlike the English SVO word order; (2) nouns take different suffixes when in a subject or object position (*at* and *ã*, respectively); (3) verbs take a suffix for 1st person singular pronomial subject or object (*li* and *sa*, respectively). Notice that, crucially, it is not possible to learn the function of the prefix *sa*, which corresponds to *me* in English, without deducing that *lhiyohli* corresponds to the verb *chases* and that third person agency in Chickasaw is not explicitly expressed. As demonstrated, inferring a translation model requires iterative reasoning on the level of words, morphemes and syntactic abstractions (classes), or, to put things differently, it requires meta-linguistic awareness.

7.3 THE DATASET

The puzzles from Linguistic Olympiads cover many aspects of language such as phonetics, morphology, syntax and semantics. They are carefully designed by experts according to several key criteria: (1) The puzzles should be *self-contained* and *unambiguous*, meaning that no prior knowledge in the foreign language is requires, just the command of one's own native language and some level of meta-linguistic awareness and that a solution is guaranteed; (2) They should require no specialized external knowledge or formal linguistic knowledge, i.e. linguistic terms are either excluded from the instructions that accompany a puzzle or they are explicitly defined; (3) The foreign language used in a puzzle should be from a truly lesser known language family (e.g. Chickasaw, Lakhota, Khmer, Ngoni), such that there is no unfair advantage to participants whose native language is related.

We based our data collection efforts on a rich and publicly available database of language puzzles maintained by the organizers of NACLO.[2] This resource contains puzzles from IOL and a wide range of local competitions[3]. We only included puzzles written in English (or translated to English) to ensure a quality transcription and to enable error analysis. Expert solutions are available for most puzzles; we excluded the rest. In addition to the translation puzzle type shown in Table 7.1, we also collected 'matching' puzzles. These are two-step puzzles, in which the participants first align a shuffled set of sentences to obtain parallel data, and then translate a set of unseen sentences. We converted these puzzles to the translation puzzle format by referring to the solution files to align the training sentence pairs. Appendix A describes how we selected the puzzles and how we transcribed them into a machine-readable format.

The final dataset contains 96 unique puzzles from 81 languages that span 39 different language families from all over the world, as well as two creoles and two artificial languages (see Appendix F for the full list). Some of the large language families have multiple representatives, e.g. there are 13 Indo-European languages, seven Austronesian and six from the Niger-Congo family. But the majority of languages are single representatives of their respective family. This genealogical diversity leads to a great diversity in the linguistic phenomena attested in the data. Some puzzles are designed to explore a specific aspect of the unknown language in isolation, e.g. case markers on demonstrative pronouns in Hungarian (Pudeyev, 2009). In general, however, the correct solution of a puzzle involves processing on the level of syntax, morphology, phonology, and semantics all at once.

[2] http://tangra.cs.yale.edu/naclobase/

[3] NACLO (North America), OzCLO (Australia), UKLO (UK), Olimpíada Brasileira (Brazil), OLE (Spain), Panini (India), Russian LO, Russian Little Bear, Swedish LO, Polish LO, Estonian LO, Slovenian LO, Bulgarian LO, Netherlands LO and more.

7.3.1 *Linguistic Phenomena*

Source	balan	waymin	bambun	baNgu	jugaNgu	jamiman.
Gloss	OBJ	mother-in-law	healthy	SUBJ	sugar-SUBJ	fat-MAKE.
Target	Sugar makes the healthy mother-in-law fat.					

Table 7.2: Example sentence in Dyibral.

The foreign languages used in linguistic puzzles are purposefully chosen to demonstrate some interesting linguistic phenomena, not found in English (or in the respective source language of the puzzle) (Bozhanov and Derzhanski, 2013), resulting in a challenging, non-trivial translation process between these diverse languages and English. In this section, we outline some key linguistic properties of the languages found in the dataset, but the list is by no means exhaustive.

SYNTAX: Three common configurations for the order between subject (S), verb (V) and object (O) in a sentence are exemplified in the dataset: SVO, SOV and VSO. In addition to these three, our dataset covers the rather rare OSV word order: see the example in Table 7.2 from the Australian language Dyirbal (Semenuks, 2012).

MORPHOLOGY: We see examples of highly analytic languages (e.g. Yoruba from West Africa) as well as highly polysythetic ones (e.g. Inuktitut from Canada). Within the synthetic type, we see both agglutinative languages (e.g. Turkish) and inflectional ones (e.g. Polish). Some specific morphological properties explored in the puzzles are verbal inflection with its many categories concerning tense, aspect and mood, nominal declension and noun class systems. The aforementioned "Dyirbal" puzzle also exemplifies an interesting classification of nouns, wherein women and dangerous animals and objects are treated as one class, men and other animals constitute another class and a third class captures all remaining nouns. The choice of the articles *balan* and *bagu* in Table 7.2, for example, is guided by this classification.

PHONOLOGY: A wide range of phonological assimilation processes interplay with the morphological processes described above and obfuscate morpheme boundaries. These can concern voicing, nasality and vowel quality, among other features. As an example of morphological and phonological processes working together, consider the realization of pronomial possession in Australian language Wembawembda (Laughren, 2009). Unlike English, which expresses this feature with pronouns *his/her/its*, Wembawemba expresses it with a suffix on the noun it modifies, e.g. *wutyupuk* '(his/her/its) stomach'. The form of

Form	nyuk	duk	nuk	buk	guk	uk
After	vowel	n	r	m	ng	other

Table 7.3: Variants of a possessive suffix in Wembawemba and their phonological distribution.

the suffix, however, depends on the ending of the noun it attaches to and can vary greatly as shown in Table 7.3.

SEMANTICS: Semantics come into play when we consider the compositionality of language and figurative speech: the phrase "falepak hawei" in the Indonesian language Abui, for example, literally translates into "pistol's ear", but a more fitting translation would be "trigger" (Peguševs, 2017).

As a side note, it is important to note that while here we use extensive linguistic terminology to discuss the properties of the languages in our dataset, the high-school students who participate in Linguistic Olympiads need not and may not be familiar with any of the terminology. Their good performance depends on a well-developed meta-linguistic awareness, not on formal linguistic training.

7.3.2 *Dataset statistics*

In total, 2311 parallel instances are transcribed—1559 training and 752 test. 63% of the test pairs are in the English → foreign direction, while the rest are in the foreign → English direction.

Statistics concerning the number of words per sentence[4] are shown on the left of Figure 7.1. The majority of both training and test pairs are fairly short, but length varies considerably. This is due to the fact that some puzzles in the dataset concern the translation of individual words, some take scope over noun-modifier phrases and some, over entire sentences. English sentences are generally longer (median 4) than their translations (median 2). This is rather intuitive considering the synthetic nature of many of the foreign languages in the dataset, wherein a single long word in the foreign language may translate into 4-5 words on the English side, as in this translation from *tΛckotoyatih* in the Mexican language Zoque to the English *only for the tooth*.

Sentence statistics about the length of the train and test split per problem are shown on the right of Figure 7.1. Intuitively, train splits are bigger than test splits. However, the number of training instances varies greatly between the puzzles, which is related to a number of factors such as the difficulty and type of the task, as well as the linguistic properties of the foreign language.

[4] We naively tokenize on space.

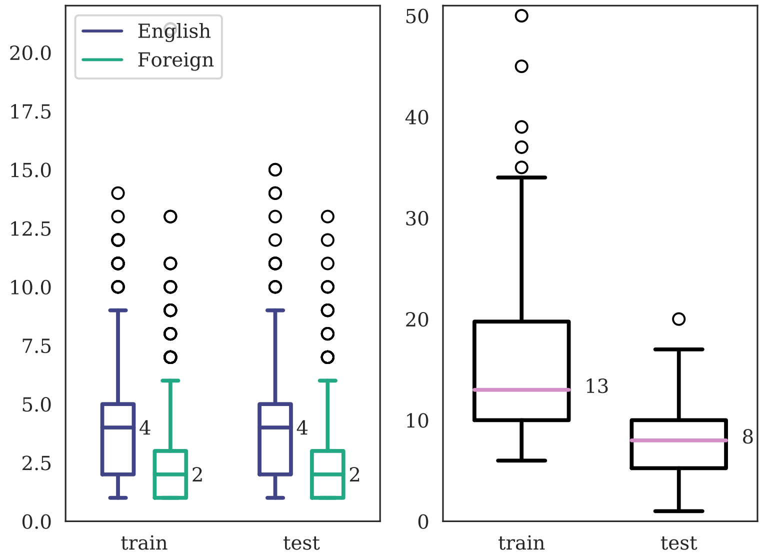


Figure 7.1: Box-plots for **Left:** Word# per language and split, **Right:** Sentence# per split.

7.3.3 *Train versus Test Splits*

One property of the data splits in linguistic puzzles, which diverges from the standard paradigm in machine learning, is that the *input* test data should not be considered "held out". On the contrary, in some cases, vocabulary items attested in the input of foreign→English test instances may be crucial to the translation of English→foreign test instances, and vice versa. So it is only the *targets* of test instances that should be truly held out. This specificity is not ubiquitous across the puzzles, but it should be accounted for by any approach to their solution, for example by building the system vocabulary over the union of the train and input test data.

7.4 BASELINES

We attemp to solve these puzzles with models of varying complexity, i.e. from random guessing to state-of-the-art neural machine translation systems.

RANDOM WORDS (RW): Since the vocabularies of source and target languages are quite small, we test what random word picking can accomplish. We simply tokenize the training sentence pairs and then randomly choose a word from the target language's vocabulary for each token in the source sentence.[5]

FASTALIGN (FA): We use the translation alignment tool FastAlign (Dyer, Chahuneau, and Smith, 2013), to test whether the puzzles can be solved by early lexical translation models (Brown et al., 1993). Since FA produces alignments for each training pair, we postprocess the output to create a translation dictionary separately for each direction. We then randomly choose from the translation entries for each token in source test sentence. [6]

PHRASE BASED STATISTICAL MACHINE TRANSLATION (PBSMT) Since Koehn and Knowles (2017) report that PBSMT models outperform vanilla NMT models in case of small parallel training data, we use PBSMT as one of the baselines. For the foreign→English direction, we implement two models—one using no external mono-lingual English data and one otherwise.

7.4.1 *Neural Machine Translation*

We implement three different models based on Transformers (Vaswani et al., 2017) using the implementation of Ott et al. (2019). In the first scenario, we train an off-the-shelf Transformer encoder-decoder model for each direction, referred to as *Transformer*. Second, we use a strong pretrained English language model, RoBERTa (Liu et al., 2019), to initialize the encoder of the NMT model for English to foreign translation. Finally, for foreign to English translation, we concatenate the translation features extracted from the last Transformer decoder layer, with the language modeling features extracted from RoBERTa (Liu et al., 2019), before mapping the vectors to the output vocabulary. These models are denoted as *Transformer+RoBERTa*.

7.5 EXPERIMENTS

7.5.1 *Experimental Settings*

We first compile a subset from the puzzles that are diverse by means of languages and contain translation questions in both directions. During tuning, we use the test sentences on these puzzles to validate our models. Since our foreign languages are morphologically rich, we use BPE (Sennrich, Haddow, and Birch, 2016) to segment words into

[5] We don't use frequency of the words, i.e., pick words that occur more often, since they are not that meaningful due to the tininess of the data.

[6] We add all aligned target phrases of the source token to the dictionary. Hence, when one target phrase is seen multiple times, it is more likely to be chosen during inference.

subwords. For the sentences in the foreign language, we learn the BPE from the training data, while for English sentences we use the already available GPT2-BPE dictionary to exploit English language prior. For convenience, before we train the models, we lowercase the sentences, remove certain punctuations, remove pronoun tags and brackets, and augment training data with multiple reference translations.

PBSMT: We use Moses (Koehn et al., 2007) with default settings. We employ wikitext-103 corpus to train a 5-gram English LM for the model with access to external data. The other model only uses training sentences for the LM.

NMT: Following the suggestions for low-resource NMT systems by Sennrich and Zhang (2019), we use small and few layers and high dropout rates. Similarly we use the smallest available language model (RoBERTa Base) and freeze its parameters during training to reduce the number of trainable parameters. We tune the following hyper-parameters: BPE merge parameter, learning rate and number of epochs.

7.5.2 *Evaluation Metrics*

The submissions to Linguistic Olympiads are manually graded by experts. For a full mark, an exact solution has to be provided, as well as a correct and detailed discussion of the underlying processes that led to this solution, e.g., concerning findings about word-order, the function of individual morphemes, etc. Participants are also given partial marks in case of partial solutions or valid discussions. Since we don't have access to expert evaluation, we use readily available automatic machine translation measures. We also note grading of system interpretations or its solution steps as an interesting future research direction.

The first is the BLEU (Papineni et al., 2002) score since it is still the standard metric in MT. We use BLEU-2 to match the lower median of sentence lengths we observe across the English and the foreign data (see Fig 7.1). BLEU matches whole words rather than word pieces, which prevents us from assigning partial credit to subword matches, which could be especially relevant for foreign target languages with rich morphology. We therefore use three additional metrics that operate on the level of word pieces: CharacTER (Wang et al., 2016), ChrF (Popovic, 2016) and ChrF++ (Popovic, 2017). CharacTER is a measure derived from TER (Translation Edit Rate), where edit rate is calculated on character level, whereas shift rate is measured on the word level. It calculates the minimum number of character edits required to adjust a hypothesis, until the reference is matched, normalized by the length of the hypothesis sentence. For easier comparison, we report $1.0 - characTER$ scores. ChrF is a simple F-measure reflecting precision and recall of the matching character n-grams. ChrF++ adds word unigrams and bi-grams to the standard ChrF for a higher human correlation score. We experiment with different combinations of character n-grams ($n = 3, 5$ as suggested in Popovic (2016)) and word n-grams ($n = 0, 1, 2$ as suggested in Popovic (2017)).

Finally, we also measure the average exact match of the puzzles, which is calculated as 1 if the prediction and reference sentences match and 0 otherwise. As it is not feasible to report and compare results on all of these metrics (nine in total), we compute the pairwise Pearson correlation coefficient between them, and average over all pairs to arrive at the following four metrics that show the least correlation with each other: BLEU−2, CharacTER, ChrF−3 and exact match. We note, however, that of these four, exact match is really the most meaningful metric. Since the sentences in the dataset are rather short and the puzzles are designed to be solvable and unambiguous, an exact match should be attainable. Moreover, as the puzzles in the dataset are of varying difficulty, the average exact match score can be seen as a continuous metric.

7.6 RESULTS AND ANALYSIS

We report the results for the best models in Fig. 7.2. The hyperparameter configuration and the development set results are given in Appendix D. The maximum exact match score among all results is only 3.4%; and the highest scores are consistently achieved by PBSMT models on both directions and dataset splits.

The overall results for foreign → English are generally higher than English → foreign. This may be due to (a) having longer sentences for English; (b) the scores (except from EM) being more suitable for English (even the character-based ones) or (c) the more challenging nature of translation into foreign languages, which needs another dedicated study.

ENGLISH→FOREIGN: Initializing the NMT encoder with RoBERTa has severely worsened the results, compared to standard Transformer model. We believe the main reason is the imbalance between encoder (huge encoder) and the decoder (tiny decoder), that makes training very challenging. The gap between the simplest baselines (RW, FA) and more sophisticated models (Transformers, PBSMT) is also considerably small; FA even surpassing Transformers's CTER and ChrF performance. For most of the foreign languages, even when two words are semantically distant, there may still be significant morpheme overlap. These suggest that simple lexical alignment models (including random assignment) can achieve higher *partial* matching scores that hints at the unreliability of CTER and ChrF measures for the puzzles.

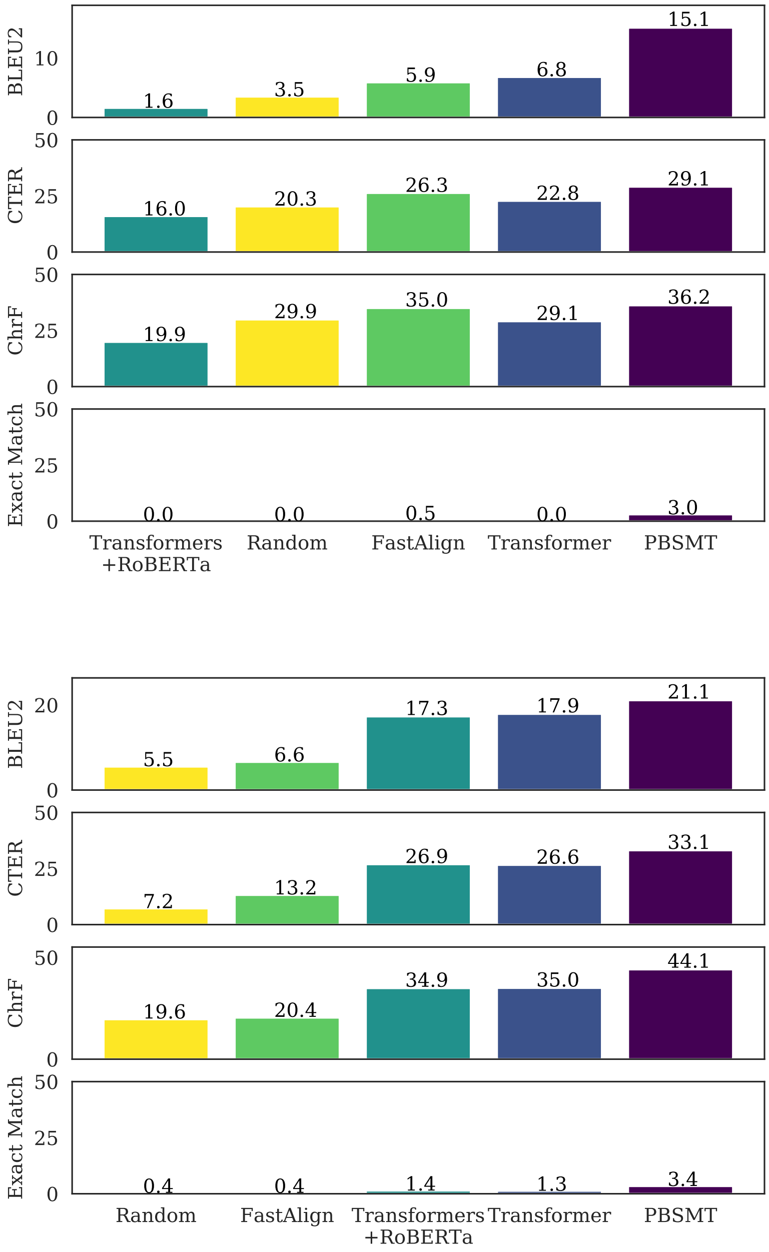


Figure 7.2: Main results (best viewed with color). **Left:** English→foreign **Right:** foreign→English.

	Chikasaw	English	PBSMT	Transformer
	Now you can translate the following into Chickasaw:			
(1)	Hattakat ihooã hollo.	The man loves the woman.	the the woman hattakã hollo	ihooat hattakã hollo
(2)	Kowi'at shoha.	The cat stinks.	the lhiyohli stinks	ofi'at shoha
(3)	Holloli.	I love her/him.	i love him	lhiyohlili
	Translate the following into English:			
(4)	Ihooat sahollo.	The woman loves me.	ihoothe sahollo	the woman loves the man
(5)	Ofi'at hilha.	The dog dances.	the(he/she) dances	the cat chases the dog
(6)	Kowi'ã lhiyohlili.	I chase the cat.	cat ch thei chase (him/her)	the dog stinks

Table 7.4: Predictions for the "Chickasaw" puzzle. Gold-standard target sentences are shown in yellow.

FOREIGN→ENGLISH: We observe that the gap between the simple and more sophisticated baselines are higher in this direction by means of all measures, as we would expect. Using RoBERTa features in the decoder does not hurt the performance while providing a small increase in EM score compared to standard Transformers. It should be noted that the decoder is still tiny and LM features are only incorporated via a separate linear layer at a very late stage, which prevents the imbalance problem we saw in English → foreign.

We see similar results for the validation data with the exception that Transformer-based models achieve either higher or the same EM scores than PBSMT while surpassing PBSMT's BLEU-2 scores in foreign → English. It supports the findings of Sennrich and Zhang (2019), drawing attention to the importance of hyper-parameter tuning for low-resource NMT models.

7.6.1 *Error Analysis*

We perform manual error analysis on the predictions of our top two models for the Chickasaw puzzle presented in Table 7.1. The predicted translations are shown in Table 7.4. We also provide the predictions of the simple baselines in Appendix E for comparison. Although the PBSMT model is best on average, we find that for this particular puzzle, the Transformer model did much better. PBSMT had very few hits overall: it correctly chose to include the lexical items *hattak* and *hollo* in (1), but the position and inflection of the former is incorrect. In (5) and (6) there are indications of correct lexicon induction, but the overall quality of the translations is very poor both in terms of accuracy and fluency. The Transformer model, on the other hand, predicts fluent translations in both directions. In the direction from English to Chickasaw, we see that the model correctly acquired the relevant morphological patterns: subjects take suffix *at*, objects take suffix *ã*, and, importantly, that first person agency is expressed through suffix *li*. The translations are still not perfect, though, due to lexical confusion: the words for *cat* and *dog* have been swapped in both (1) and (2), as well as the words for *love* and *chase* in (3). In the direction from Chickasaw to English, the Transformer's predictions remain fluent, but they hardly relate to the input. Contrary to the overall results, for this puzzle translation *to* English appears to be more challenging for the model.

7.7 RELATED WORK

Recently, reasoning tasks and datasets that require natural language processing have been introduced, such as common-sense reasoning in the form of pronoun resolution e.g., WSC (Levesque, 2011), multiple-choice question answering e.g., SWAG (Zellers et al., 2018) and ARC (Clark et al., 2018); inference tasks in the form of binary or multi-label classification problems e.g., the GLUE benchmarks (Wang et al., 2018); and visual reasoning in the form of question answering (Zellers et al., 2019) e.g., GQA (Hudson and Manning, 2019a). In these tasks, the required level of semantics is mostly limited to single sentences rather than a collection; almost all tasks target English; data is derived from running text and is mostly close-domain. In addition, some require external knowledge bases or high-level knowledge on physical models or experiments as in ARC classified by Boratko et al. (2018), which leaves room for accumulating errors from external parts and complicates the analysis of individual parts like reasoning.

Another body of early work on symbolic AI provides a different set of tools to model reasoning such as rule-engines, rule-induction algorithms, logic programs and case-based reasoning models (Kolodner, 1992). However, it is not trivial to represent and model our task in these frameworks, since they mostly require defining primitives, expressions, discrete features and cases. Furthermore, the strength of statistical/neural models has been repeatedly shown to surpass rule-based models. Our goal is to encourage researchers to incorporate reasoning into statistical models, rather than replacing them with symbolic models.

7.8 CONCLUSION AND FUTURE WORK

The field of NLP has developed deep neural models that can exploit large amounts of data to achieve high scores on downstream tasks. Still, the field lacks models that can perform human-like reasoning and generalization. To mitigate this gap, we draw inspiration from the *Linguistic Olympiads* that challenge the meta-linguistic and reasoning abilities of high-school students. We create a new benchmark dataset from available Linguistic Puzzles that spans over 81 languages from 39 language families, which is released at `https://ukplab.github.io/PuzzLing-Machines/`. We implement and evaluate simple baselines such as alignment, and state-of-the-art machine translation models with integrated a pretrained English language model. We show that none of the models can perform well on the puzzles, suggesting that we are still far from having systems with meta-linguistic awareness and reasoning capabilities.

acknowledgements

This work was supported by the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1 and grant GU 798/17-1). We would like to thank Liane Vogel, Marc Simon Uecker and Siddharth Singh Parihar for their great help during the project. We are grateful to Dragomir Radev for his feedback and continuous help with encoding problems encountered during puzzle transcription. We thank Adam Lopez and Ilia Kuznetsov for providing feedback on early drafts of the paper. We thank the area chairs and the senior area chair, whose comments helped us improve the paper.

appendices

A Transcription of Puzzles

The puzzles are generally provided as pdf files. Many languages in the dataset use the Latin script, optionally with some diacritics. Some which use a non-Latin script (or have no writing system at all), are transcribed with IPA or transliterated into the Latin script. Only one language, Georgian, uses a non-Latin script, namely the Mkhedruli script. As there are various types of puzzles presented at the Olympiads, we identified the ones relevant to our task through automatic filtering for the keywords "translation" or "matching", and manually verified the results.

To represent linguistic puzzles in a unified, machine-readable format, we defined a JSON format shown in Appendix B. The relevant data was manually extracted from the PDF files and mapped to this format in a semi-automated fashion. We faced encoding issues with many of the puzzles. For some of these, the database owner kindly provided us with the source files of the pdf documents, which enabled us to generate UTF-8 encoding of the data; others we fixed manually. Some puzzles, which use pictorial scripts or are otherwise UTF-8 incompatible, were discarded.

During the transcription we came across various formats of linguistic annotation in the puzzles. This kind of information was not consistently provided across puzzles, but we included it where available, as it can be both helpful and crucial to a correct solution. In the next paragraphs, we provide details on the different types of annotated information and the standardized format we used to encode those.

	Language	Source sentence	Target sentence	Other accepted forms
1.	Chickasaw	Hilha.	(She/He) dances.	She dances. He dances.
2a.	Blackfoot	Nitoki'kaahpinnaan.	We.PL2- camped.	We camped.
2b.	Blackfoot	Oki'kaao'pa.	We.PL2 camped.	We camped.
3.	Wambaya	Bardbi ga bungmanya.	The old woman ran [away].	The old woman ran away.
4.	Euskara	Umea etorri da.	The child has (come/arrived).	The child has come. The child has arrived.

Table 7.5: Examples of special transcription notation.

GENDER DISTINCTION IN PRONOUNS: When the foreign language does not mark gender on pronouns (or omits pronouns altogether), singular pronouns in the English translations are provided consistently as *(he/she)* and *(him/her)*, or *(he/she/it)* and *(his/her/its)*, as in Ex. 1 in Table 7.5. During evaluation, instances of this notation are accepted, as well as instances of the individual alternatives.

NUMBER MARKING ON PRONOUNS: When the foreign language marks two levels of plurality for the second person pronoun *you*, they are marked accordingly as *you.SG* and *you.PL*. Some languages make a distinction between plural forms concerning two objects and plural forms concerning three or more objects. In this case, we mark pronouns (not just *you*, but also *we* and *they*) with the notation *.PL2* and *.PL3*, respectively. Some languages also make a distinction between an inclusive *we* 'you and me' and an exclusive *we* 'me and someone else'. We reserve *we.PL2* for the inclusive sense, and mark the exclusive sense with *we.PL2-*. See examples 2a and 2b in Table 7.5. The notation presented here holds for both personal pronouns, e.g. *you*, and possessive pronouns, e.g. *your*. During evaluation, we disregard this notation on the side of the target language.

ZERO-TO-ONE MATCHING: Words that are semantically implied or required by English grammar, but not directly expressed on the side of the foreign language are shown in square brackets in some of the puzzles, as in Table 7.5-Ex. 3. This bracketing exists only to aid the learning of a translation model. During evaluation, we remove these brackets from the target test sentences.

Notice that number marking and special notation for zero-to-one matching is not ubiquitous across the puzzles. We included it only when it was provided in the original puzzle.

MULTIPLE REFERENCE TRANSLATIONS: Occasionally, several possible translations are listed in a puzzle for a given word, phrase or sentence–see Table 7.5-Ex. 4. We represent these options inside parenthesis separated with a slash (/), e.g., (alternative1/.../alternative N). Since the alternatives are of different granularity, nested bracketing may sometimes occur. During evaluation, we calculate the scores between the prediction and all possible references, and take the maximum.

ADDITIONAL INFORMATION Roughly half of the puzzles contain remarks on individual characters and diacritics in the inventory of the foreign language, e.g. "In the following orthography a colon (:) marks a long vowel, and the ʔ symbol marks a glottal stop (like the sound in the middle of uh-oh)". In many cases, the instructions state that these are pronunciation notes, i.e. they are made available only to allow the participants to vocalize the words they see on the page. On some occasions, however, they might introduce a character that is not present in the training data, but is relevant to the translation of the test sentences, e.g. the voiceless counterpart of a voiced consonant in a language with voice assimilation. As this kind of information cannot be mapped to the parallel data format, we include it in a separate field in the JSON files, directly as it appeared in the puzzles. [7]

With the aforementioned guidelines, each puzzle was transcribed by one transcriber and verified by at least one other transcriber. For the test pairs, the direction of translation is stored as well, since a possible and singular solution is only guaranteed in the direction as given in the puzzle.

B JSON File Format

Each puzzle is represented with a JSON file containing the following fields: SOURCE_LANG, TARGET_LANG, META, TRAIN and TEST. Each field is explained in Table 7.6.

C Development Results

The results on the validation set are given in Fig. 7.3.

D Hyperparameter Settings

The best hyperparameters found for each NMT model is given as following. *FA:* word to word alignments; PBSMT for English→Foreign: word alignment with external English LM; PBSMT for Foreign→English: BPE with 30 merge operations. For both Transformers-based models in Foreign→English direction, we used BPE with 10 merge operations, learning rate of 0.001 and 500 epochs; while for the standard Transformer in English→Foreign direction, BPE with 30 merge operations have been used. For all models except from Transformers with RoBERTa encoder, both the encoder and decoder had 1 layers,

[7] We believe that even if all instances of such remarks are ignored, the puzzles should remain mostly solvable, but we note that without this information, the ceiling for human performance would not be quite 100 percent.

Field	Definition	Example
SOURCE_LANG	Name of the source language	Foreign language e.g., Kiswahili, Pali
TARGET_LANG	Name of the target language	English
META	Additional information about the foreign language provided in the original puzzle (as free text)	"The sound represented as ã is a ’nasalized’ vowel. It is pronounced like the ’a’ in ’father’, but with the air passing through the nose, as in the French word ’ban’."
TRAIN	Parallel training sentences given as a list of lists	[[“Bonjour”, “Good morning”], [“chat”, “cat”]], where the source and the target language is French and English respectively.
TEST	Parallel test sentences with direction information	[[“Bonjour”, “Good morning”, >], [“chat”, “cat”, <]]. “>” implies that the translation is required from source to target language, vice versa for “<”

Table 7.6: JSON file format used in the linguistic puzzles shared task

Chikasaw	English	RW	FA
Hattakat ihooã hollo.	The man loves the woman.	Ihooat lhiyohli hollo salhiyohli ofi’at.	The hollo loves the woman.
Kowi’at shoha.	The cat stinks.	Lhiyohlili lhiyohlili kowi’ã.	The lhiyohli shoha.
Holloli.	I love her/him.	Ofi’ã hilha lhiyohlili.	I love lhiyohlili.
Ihooat sahollo.	The woman loves me.	Dog loves	Ihooat sahollo
Ofi’at hilha.	The dog dances.	I the	ofi’at he dances
Kowi’ã lhiyohlili.	I chase the cat.	stinks cat	Kowi’ã I chase (him/her).

Table 7.7: Predictions of the simple baseline models for the “Chickasaw” puzzle. Gold-standard target sentences are shown in yellow.

and all hidden dimesions were set to 128, dropout was set to 0.3, and the models were trained with Adam optimizer. For Transformer with RoBERTA LM Encoder for English→Foreign, we have used 0.0001 learning rate with reduction on plateau, batches of size 2, dropout of 0.1, 1 layer, 64 embedding units, 128 hidden units, and BPE with 5 merge operations.

E Chickasaw Additional Predictions

In Table 7.7, the predictions of RW and FA are shown for comparison.

F List of Languages and Families

The full list for the languages and the families they belong to, as classified in WALS (Dryer and Haspelmath, 2013) and, where WALS

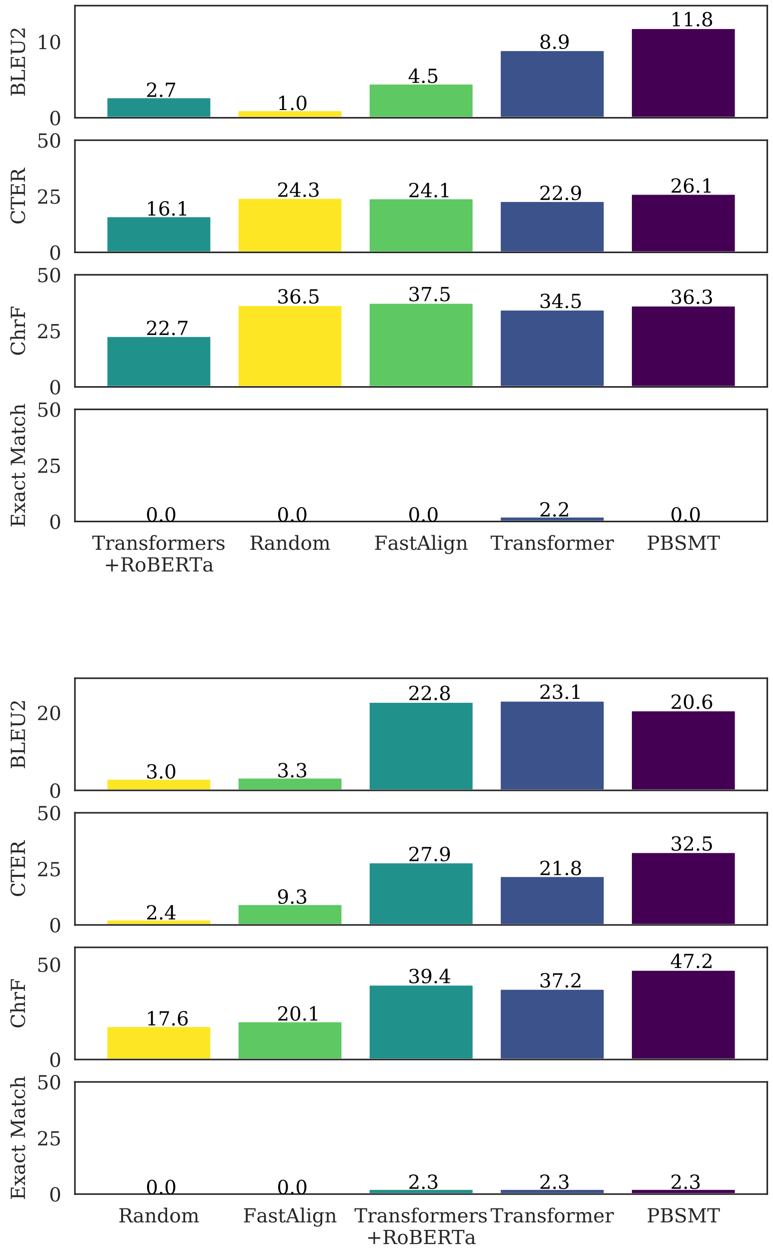


Figure 7.3: Development set results. **Left:** English→foreign **Right:** foreign→English

lacks an entry, Glottolog (Hammarström, Forkel, and Haspelmath, 2019), are given in Table 7.8.

Language	Family	Language	Family
Abkhaz	Northwest Caucasian	Luiseño	Uto-Aztecan
Abma	Austronesian	Madak	Austronesian
Abui	Timor-Alor-Pantar	Malay	Austronesian
Afrihili	Artificial	Maori	Austronesian
Amele	Trans-New Guinea	Mayangna	Misumalpan
Ancient Greek	Indo-European	Miwoc	Penutian
Bambara	Mande	Muna	Austronesian
Basque	Basque	Nahuatl	Uto-Aztecan
Beja	Afro-Asiatic	Ndebele	Niger-Congo
Benabena	Trans-New Guinea	Nen	Trans-New Guinea
Blackfoot	Algic	Nepali	Indo-European
Bulgarian	Indo-European	Nhanda	Pama-Nyungan
Central Cagayan Agta	Austronesian	Norwegian	Indo-European
Chamalal	Nakh-Daghestanian	Nung	Tai-Kadai
Chickasaw	Muskogean	Old English	Indo-European
Choctaw	Muskogean	Pali	Indo-European
Cupeño	Uto-Aztecan	Papiamento	creole
Danish	Indo-European	Persian	Indo-European
Dyirbal	Pama-Nyungan	Polish	Indo-European
Esperanto	Artificial	Proto-Algoquian	Algic
Fula	Niger-Congo	Quechua	Quechuan
Georgian	Kartvelian	Somali	Afro-Asiatic
Guaraní	Tupian	Swahili	Niger-Congo
Haitian Creole	Creole	Tadaksahak	Songhay
Hmong	Hmong-Mien	Tanna	Austronesian
Hungarian	Uralic	Teop	Austronesian
Icelandic	Indo-European	Tok Pisin	creole
Ilokano	Austronesian	Tshiluba	Niger-Congo
Inuktitut	Eskimo-Aleut	Turkish	Altaic
Irish	Indo-European	Udihe	Altaic
Jaqaru	Aymaran	Waanyi	Garrwan
Kabardian	Northwest Caucasian	Wambaya	Mirndi
Kayapo	Macro-Ge	Warlpiri	Pama-Nyungan
Kimbundu	Niger-Congo	Welsh	Indo-European
Kunuz Nubian	Eastern Sudanic	Wembawemba	Pama-Nyungan
Kurdish	Indo-European	Witsuwit'en	Dené–Yeniseian
Lakhota	Siouan	Yidiny	Pama-Nyungan
Lalana Chinantec	Oto-Manguean	Yolmo	Sino-Tibetan
Latvian	Indo-European	Yonggom	Nuclear Trans New Guinea
Lopit	Nilo-Saharan	Yoruba	Niger-Congo
		Zoque	Mixe-Zoque

Table 7.8: Full list of languages and their families.

8 CONCLUSION

The work presented in this thesis contributes to efforts in the field of low-resource, multilingual NLP. It does so through novel methods, evaluations and resources, which provide insights into the existing multilingual NLP solutions and avenues for further research.

From Chapters 2, 3 and 4 of the thesis we can conclude that while apparent improvements can be achieved in CLWE alignment, both supervised and unsupervised, it is important to evaluate those fairly and with recourse to the flaws of the evaluation data available. Even in the era of large pretrained transformer-based language encoders, which have shown promising results in the space of cross-lingual transfer (Wu and Dredze, 2019), cross-lingual word embedding alignment still finds its applications (Cao, Kitaev, and Klein, 2020; Schuster et al., 2019). Therefore the findings presented here remain relevant.

In Chapter 5, an analysis of the source of performance in dependency parsing models trained with cross-lingual transfer showed that even in the absence of high vocabulary overlap between source and target language, transfer of knowledge on the level of syntax (part of speech) from the source language can lead to improved dependency parsing of the target language. This finding corroborates more recent work on the transfer of knowledge through multilingual transformer-based language encoders showing that vocabulary overlap is not required for transfer of knowledge to occur (Artetxe, Ruder, and Yogatama, 2020).

Chapter 6 showed that zero-shot cross-domain transfer learning for appositive generation is not sufficiently robust across the four languages explored. Improvements over this baseline can be expected to come from the application of domain adaptation methods (Jiang, 2008), as well as more elaborate methods for querying a knowledge base.

In Chapter 7, we showed that no current NLP model is equipped to solve the PuzzLing challenge dataset, and argued that advancements in the space of meta-learning (Finn, Abbeel, and Levine, 2017), along with multi-hop reasoning and abstract learning (Hudson and Manning, 2019b), will bring NLP models to the level of sample-efficiency and reasoning skill needed to solve a complex task from just a few examples.

Low-resource language scenarios certainly pose a challenge in NLP, but advancements in this area are an important step towards the democratization of the Internet and of language technology. The lack of high-quality language technology for these languages is currently setting them even further back in terms of socioeconomic status, whereas the introduction of new technology that supports them has the potential to reverse this process.

BIBLIOGRAPHY

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg (2017). „Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks." In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Agić, Željko (2017). „Cross-Lingual Parser Selection for Low-Resource Languages." In: *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*. Gothenburg, Sweden, pp. 1–10.

Alvarez-Melis, David and Tommi Jaakkola (2018). „Gromov-Wasserstein Alignment of Word Embedding Spaces." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 1881–1890.

Ammar, Waleed, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith (2016a). „Many Languages, One Parser." In: *Transactions of the Association for Computational Linguistics* 4, pp. 431–444.

Ammar, Waleed, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith (2016b). „Massively multilingual word embeddings." In: *arXiv preprint arXiv:1602.01925*.

Angeli, Gabor, Percy Liang, and Dan Klein (Oct. 2010). „A Simple Domain-Independent Probabilistic Approach to Generation." In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA, pp. 502–512.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). „Wasserstein GAN." In: *CoRR*, abs/1701.07875.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2017). „Learning bilingual word embeddings with (almost) no bilingual data." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 451–462.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2018a). „A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, pp. 789–798.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2018b). „Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations." In: *Proceedings of AAAI 2018*.

Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama (July 2020). „On the Cross-lingual Transferability of Monolingual Representations." In: *Proceedings of the 58th Annual Meeting of the Association for Computa-*

tional Linguistics. Online: Association for Computational Linguistics, pp. 4623–4637.

Balaraman, Vevake, Simon Razniewski, and Werner Nutt (2018). „Recoin: relative completeness in Wikidata." In: *Companion Proceedings of the The Web Conference 2018*, pp. 1787–1792.

Balles, Lukas, Javier Romero, and Philipp Hennig (2017). „Coupling Adaptive Batch Sizes with Learning Rates." In: *Proceedings of UAI*.

Barone, Antonio Valerio Miceli (2016). „Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders." In: *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 121–126.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta (2009). „The WaCky wide web: a collection of very large linguistically processed web-crawled corpora." In: *Language resources and evaluation* 43.3, pp. 209–226.

Bauer, B.L.M. (2017). *Nominal Apposition in Indo-European: Its Forms and Functions, and its Evolution in Latin-Romance*. Trends in Linguistics. Studies and Monographs [TiLSM]. De Gruyter.

Bengio, Yoshua (Feb. 2020). *Deep Learning for AI.* Invited talk at AAAI.

Bergmanis, Toms, Katharina Kann, Hinrich Schütze, and Sharon Goldwater (2017). „Training Data Augmentation for Low-Resource Morphological Inflection." In: *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*. Vancouver: Association for Computational Linguistics, pp. 31–39.

Besl, Paul and Neil McKay (1992). „A Method for Registration of 3-D Shapes." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14.2.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). „Enriching Word Vectors with Subword Information." In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Boratko, Michael et al. (2018). „A Systematic Classification of Knowledge, Reasoning, and Context within the ARC Dataset." In: *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*, pp. 60–70.

Bozhanov, Bozhidar and Ivan Derzhanski (2013). „Rosetta Stone Linguistic Problems." In: *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pp. 1–8.

Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993). „The Mathematics of Statistical Machine Translation: Parameter Estimation." In: *Computational Linguistics* 19.2, pp. 263–311.

Brutzkus, Alon, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz (2018). „SGD Learns Over-parameterized Networks that Provably Generalize on Linearly Separable Data." In: *Proceedings of ICLR*.

Cao, Steven, Nikita Kitaev, and Dan Klein (2020). „Multilingual alignment of contextual word representations." In: *Proceedings of ICLR*.

Chomsky, Noam A (1976). *Reflections on language*. New York: Pantheon.

Clark, Peter, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord (2018). „Think you have solved question answering? Try ARC, the AI2 reasoning challenge." In: *arXiv preprint arXiv:1803.05457*.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou (2018). „Word Translation Without Parallel Data." In: *Proceedings of ICLR 2018*.

Czarnowska, Paula, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake (Nov. 2019). „Don't Forget the Long Tail! A Comprehensive Analysis of Morphological Generalization in Bilingual Lexicon Induction." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 974–983.

Daskalakis, Constantinos, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng (2018). „Training GANs with Optimism." In: *Proceedings of ICLR 2018*.

Denkowski, Michael and Alon Lavie (June 2014). „Meteor Universal: Language Specific Translation Evaluation for Any Target Language." In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 376–380.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *arXiv preprint arXiv:1810.04805*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

Dinu, Georgiana, Angeliki Lazaridou, and Marco Baroni (2015). „Improving Zero-Shot Learning by Mitigating the Hubness Problem." In: *Proceedings of ICLR 2015 (Workshop Track)*.

Doval, Yerai, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert (2018). „Improving Cross-Lingual Word Embeddings by Meeting in the Middle." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 294–304.

Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith (June 2013). „A Simple, Fast, and Effective Reparameterization of IBM Model 2." In:

Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, pp. 644–648.

Ezra, Esther, Micha Sharir, and Alon Efrat (2006). „On the ICP Algorithm.“ In: SCG 06. Sedona, Arizona, USA: Association for Computing Machinery, pp. 95–104.

Fadaee, Marzieh, Arianna Bisazza, and Christof Monz (2017). „Data Augmentation for Low-Resource Neural Machine Translation.“ In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada, pp. 567–573.

Falenska, Agnieszka and Özlem Çetinoğlu (2017). „Lexicalized vs. Delexicalized Parsing in Low-Resource Scenarios.“ In: *Proceedings of the 15th International Conference on Parsing Technologies*. Pisa, Italy: Association for Computational Linguistics, pp. 18–24.

Faruqui, Manaal and Chris Dyer (2014). „Improving Vector Space Word Representations Using Multilingual Correlation.“ In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 462 –471.

Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer (Aug. 2016). „Problems With Evaluation of Word Embeddings Using Word Similarity Tasks.“ In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany, pp. 30–35.

Finn, Chelsea, Pieter Abbeel, and Sergey Levine (2017). „Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks.“ In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, pp. 1126–1135.

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer (July 2018). „AllenNLP: A Deep Semantic Natural Language Processing Platform.“ In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Melbourne, Australia, pp. 1–6.

Gerz, Daniela, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen (Nov. 2016). „SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity.“ In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, pp. 2173–2182.

Glavaš, Goran, Robert Litschko, Sebastian Ruder, and Ivan Vulić (July 2019). „How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions.“ In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 710–721.

Gold, S and A Rangarajan (1996). „A graduated assignment algorithm for graph matching." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18, pp. 377–388.

Goodfellow, Ian J, Oriol Vinyals, and Andrew Saxe (2015). „Qualitatively characterizing neural network optimization problems." In: *Proceedings of ICLR 2015*.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). „Generative Adversarial Nets." In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger. Vol. 27. Curran Associates, Inc.

Gouws, Stephan and Anders Søgaard (2015). „Simple task-specific bilingual word embeddings." In: *Proceedings of NAACL-HLT*, pp. 1302–1306.

Gower, John C (1975). „Generalized procrustes analysis." In: *Psychometrika* 40.1, pp. 33–51.

Grave, Edouard, Armand Joulin, and Quentin Berthet (2019). „Unsupervised Alignment of Embeddings with Wasserstein Procrustes." In: Proceedings of Machine Learning Research 89. Ed. by Kamalika Chaudhuri and Masashi Sugiyama, pp. 1880–1890.

Gu, Jiatao, Zhengdong Lu, Hang Li, and Victor O.K. Li (Aug. 2016). „Incorporating Copying Mechanism in Sequence-to-Sequence Learning." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, pp. 1631–1640.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni (2018). „Colorless Green Recurrent Networks Dream Hierarchically." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana, pp. 1195–1205.

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville (2017). „Improved Training of Wasserstein GANs." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith (2018). „Annotation Artifacts in Natural Language Inference Data." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pp. 107–112.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath (2019). *Glottolog 4.1*. Max Planck Institute for the Science of Human History. Jena.

Hauer, Bradley, Garrett Nicolai, and Grzegorz Kondrak (Apr. 2017). „Bootstrapping Unsupervised Bilingual Lexicon Induction.“ In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain, pp. 619–624.

Hoshen, Yedid and Lior Wolf (2018a). „An Iterative Closest Point Method for Unsupervised Word Translation.“ In: *CoRR* abs/1801.06126.

Hoshen, Yedid and Lior Wolf (2018b). „Non-Adversarial Unsupervised Word Translation.“ In: pp. 469–478.

Howard, Jeremy and Sebastian Ruder (2018). „Universal Language Model Fine-tuning for Text Classification.“ In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 328–339.

Hudson, Drew A. and Christopher D. Manning (2019a). „GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering.“ In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709.

Hudson, Drew and Christopher D Manning (2019b). „Learning by Abstraction: The Neural State Machine.“ In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak (Sept. 2005). „Bootstrapping Parsers via Syntactic Projection Across Parallel Texts.“ In: *Nat. Lang. Eng.* 11.3, pp. 311–325.

Jastrzebski, Stanislaw, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey (2018). „Finding Flatter Minima with SGD.“ In: *Proceedings of ICLR*.

Jawanpuria, Pratik, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra (2019). „Learning Multilingual Word Embeddings in Latent Metric Space: A Geometric Approach.“ In: *Transactions of the Association for Computational Linguistics* 7, pp. 107–120.

Jiang, Jing (2008). „A literature survey on domain adaptation of statistical classifiers.“ In: *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey* 3, pp. 1–12.

Joulin, Armand, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave (2018a). „Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion.“ In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Joulin, Armand, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave (2018b). „Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion.“ In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 2979–2984.

Kahneman, Daniel (2011). *Thinking, fast and slow*. Macmillan.

Kang, Jun Seok, Robert L Logan IV, Zewei Chu, Yang Chen, Dheeru Dua, Kevin Gimpel, Sameer Singh, and Niranjan Balasubramanian (2019). „PoMo: Generating Entity-Specific Post-Modifiers in Context." In: *Proceedings of NAACL-HLT*, pp. 826–838.

Kementchedjhieva, Yova, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard (2018). „Generalizing procrustes analysis for better bilingual dictionary induction." In: *arXiv preprint arXiv:1809.00064*.

Kim, Joohyun and Raymond Mooney (Aug. 2010). „Generative Alignment and Semantic Parsing for Learning from Ambiguous Supervision." In: *COLING 2010: Posters*. Beijing, China: Coling 2010 Organizing Committee, pp. 543–551.

Kiperwasser, Eliyahu and Yoav Goldberg (2016). „Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations." In: *Transactions of the Association for Computational Linguistics* 4, pp. 313–327.

Kirov, Christo et al. (May 2018). „UniMorph 2.0: Universal Morphology." In:

Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai (2012). „Inducing Crosslingual Distributed Representations of Words." In: *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, pp. 1459–1474.

Koehn, Philipp and Rebecca Knowles (Aug. 2017). „Six Challenges for Neural Machine Translation." In: *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver, pp. 28–39.

Koehn, Philipp et al. (2007). „Moses: Open Source Toolkit for Statistical Machine Translation." In: *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. ACL '07. Prague, Czech Republic: Association for Computational Linguistics, pp. 177–180.

Kolodner, Janet L. (1992). „An introduction to case-based reasoning." In: *Artif. Intell. Rev.* 6.1, pp. 3–34.

Konstas, Ioannis and Mirella Lapata (2013). „A Global Model for Concept-to-Text Generation." In: *J. Artif. Intell. Res.* 48, pp. 305–346.

Kuhlmann, Marco, Carlos Gómez-Rodrìguez, and Giorgio Satta (2011). „Dynamic Programming Algorithms for Transition-Based Dependency Parsers." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA, pp. 673–682.

Lample, Guillaume, Ludovic Denoyer, and Marc'Aurelio Ranzato (2018). „Unsupervised Machine Translation Using Monolingual Corpora Only." In: *Proceedings of ICLR (Conference Papers)*.

Laughren, Mary (2009). *Wembawemba expressing possession*. `http://cxielamiko.narod.ru/zadachi/ozclo-09-state.pdf`, OzCLO.

LeCun, Yann (Feb. 2020). *Self-Supervised Learning*. Invited talk at AAAI.

Levesque, Hector J. (2011). „The Winograd Schema Challenge.“ In: *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*.

Lhoneux, Miryam de, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard (2018). „Parameter sharing between dependency parsers for related languages.“ In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4992–4997.

Lhoneux, Miryam de, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre (2017). „From Raw Text to Universal Dependencies – Look, No Tags!“ In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* Vancouver, Canada.

Lhoneux, Miryam de, Sara Stymne, and Joakim Nivre (2017). „Arc-Hybrid Non-Projective Dependency Parsing with a Static-Dynamic Oracle.“ In: *Proceedings of the The 15th International Conference on Parsing Technologies (IWPT).* Pisa, Italy.

Li, Hao, Zheng Xu, Gavin Taylor, and Tom Goldstein (2018). „Visualizing the Loss Landscape of Neural Nets.“ In: *Proceedings of ICLR*.

Liang, Percy, Michael Jordan, and Dan Klein (Aug. 2009). „Learning Semantic Correspondences with Less Supervision.“ In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, pp. 91–99.

Lim, KyungTae, Niko Partanen, and Thierry Poibeau (2018). „Multilingual Dependency Parsing for Low-Resource Languages: Case Studies on North Saami and Komi-Zyrian.“ In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).* Miyazaki, Japan.

Ling, Wang, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis (2015). „Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation.“ In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pp. 1520–1530.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). „RoBERTa: A Robustly Optimized BERT Pretraining Approach.“ In: *CoRR* abs/1907.11692.

Lu, Ang, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu (2015). „Deep Multilingual Correlation for Improved Word Embeddings.“ In: *HLT-NAACL*.

Makazhanov, Aibek, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev (2015). „Syntactic Annotation of Kazakh:

Following the Universal Dependencies Guidelines. A report." In: *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pp. 338–350.

Marcus, Gary (2020). „The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence." In: *CoRR* abs/2002.06177.

McClelland, James L., Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze (2019). „Extending Machine Language Models toward Human-Level Language Understanding." In: *CoRR* abs/1912.05877.

McDonald, Ryan, Slav Petrov, and Keith Hall (2011). „Multi-Source Transfer of Delexicalized Dependency Parsers." In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, pp. 62–72.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever (2013). „Exploiting Similarities among Languages for Machine Translation." In: *CoRR* abs/1309.4168.

Mimno, David and Laure Thompson (Sept. 2017). „The strange geometry of skip-gram with negative sampling." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2873–2878.

Mohammad, Saif M., Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney (2013). „Computing Lexical Contrast." In: *Computational Linguistics* 39.3, pp. 555–590.

Nivre, Joakim (2009). „Non-Projective Dependency Parsing in Expected Linear Time." In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore, pp. 351–359.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman (May 2020). „Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection." English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France, pp. 4034–4043.

Nivre, Joakim et al. (2016). „Universal Dependencies v1: A Multilingual Treebank Collection." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Portorož, Slovenia.

Nivre, Joakim et al. (2018). *Universal Dependencies 2.2*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli (2019). „fairseq: A Fast, Extensible Toolkit for Sequence Modeling." In: *Proceedings of NAACL-HLT 2019: Demonstrations*.

Pan, S. J. and Q. Yang (2010). „A Survey on Transfer Learning.“ In: *IEEE Transactions on Knowledge and Data Engineering* 22.10, pp. 1345–1359.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (July 2002). „BLEU: a Method for Automatic Evaluation of Machine Translation.“ In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318.

Payne, Tom (2005). *Chickasaw*. `http://lingclub.mycpanel.princeton.edu/challenge/chickasaw.php`, Linguistics Society of America.

Peguševs, Aleksejs (2017). *Abui*. `https://ioling.org/booklets/iol-2017-indiv-prob.en.pdf`, IOL.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (June 2018). „Deep Contextualized Word Representations.“ In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana, pp. 2227–2237.

Pęzik, Piotr (2016). „Exploring phraseological equivalence with Paralela.“ In:

Pierini, Patrizia (Jan. 2008). „Opening a Pandora's Box: Proper Names in English Phraseology.“ In: *Linguistik Online* 36.

Poincaré, Henri (1902). *La Science et 1′ Hypothese*. Paris, France: Flammario.

Popovic, Maja (2016). „chrF deconstructed: beta parameters and n-gram weights.“ In: *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pp. 499–504.

Popovic, Maja (2017). „chrF++: words helping character n-grams.“ In: *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pp. 612–618.

Pudeyev, Victor (2009). *Hungarian*. `http://tangra.cs.yale.edu/naclobase/get.cgi?pid=198&version=1`, NACLO.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning (2020). „Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.“ In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Rademaker, Alexandre, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva (2017). „Universal Dependencies for Portuguese.“ In: *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*. Pisa, Italy, pp. 197–206.

Radovanović, Milos, Alexandros Nanopoulos, and Mirjana Ivanovic (2010). „Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data.“ In: *Journal of Machine Learning Research* 11, pp. 2487–2531.

Raiko, Tapani, Harri Valpola, and Yann LeCun (2012). „Deep Learning Made Easier by Linear Transformations in Perceptrons." In: *AISTATS*.

Rasooli, Mohammad Sadegh and Michael Collins (2017). „Cross-Lingual Syntactic Transfer with Limited Resources." In: *Transactions of the Association for Computational Linguistics* 5, pp. 279–293.

Rosa, Rudolf and David Mareček (2018). „CUNI x-ling: Parsing Under-Resourced Languages in CoNLL 2018 UD Shared Task." In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium: Association for Computational Linguistics, pp. 187–196.

Rothe, Sascha, Shashi Narayan, and Aliaksei Severyn (2020). „Leveraging Pre-trained Checkpoints for Sequence Generation Tasks." In: *Transactions of the Association for Computational Linguistics* 8, pp. 264–280.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard (2018). „A Survey of Cross-lingual Word Embedding Models." In: *Journal of Artificial Intelligence Research*.

Sahin, Gozde Gul and Mark Steedman (2018). „Data Augmentation via Dependency Tree Morphing for Low-Resource Languages." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 5004–5009.

Schönemann, Peter H (1966). „A generalized solution of the orthogonal Procrustes problem." In: *Psychometrika* 31.1, pp. 1–10.

Schuster, Tal, Ori Ram, Regina Barzilay, and Amir Globerson (June 2019). „Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1599–1613.

Semenuks, Arthur (2012). *Dyirbal*. `https://ioling.org/booklets/iol-2012-indiv-prob.en.pdf`, International Linguistic Olympiad.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). „Neural Machine Translation of Rare Words with Subword Units." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725.

Sennrich, Rico and Biao Zhang (2019). „Revisiting Low-Resource Neural Machine Translation: A Case Study." In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 211–221.

Sheyanova, Mariya and Francis M. Tyers (2017). „Annotation schemes in North Sámi dependency parsing." In: *Proceedings of the 3rd Inter-*

national Workshop for Computational Linguistics of Uralic Languages, pp. 66–75.

Shi, Xing, Inkit Padhi, and Kevin Knight (2016). „Does String-Based Neural MT Learn Source Syntax?" In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1526–1534.

Smith, Aaron, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne (2018). „82 Treebanks, 34 Models: Universal Dependency Parsing with Multi-Treebank Models." In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium, pp. 113–123.

Smith, Samuel L., David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla (2017). „Bilingual word vectors, orthogonal transformations and the inverted softmax." In: *Proceedings of ICLR*.

Smith, Samuel and Quoc Le (2018). „A Bayesian perspective on generalization and stochastic gradient descent." In: *Proceedings of ICLR*.

Søgaard, Anders (2011). „Data point selection for cross-language adaptation of dependency parsers." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA, pp. 682–686.

Søgaard, Anders, Sebastian Ruder, and Ivan Vulić (July 2018). „On the Limitations of Unsupervised Bilingual Dictionary Induction." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, pp. 778–788.

Straka, Milan (2018). „UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task." In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium, pp. 197–207.

Straka, Milan and Jana Straková (2017). „Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe." In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada, pp. 88–99.

Stymne, Sara, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre (2018). „Parser Training with Heterogeneous Treebanks." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia, pp. 619–625.

Tiedemann, Jörg and Željko Agic (Jan. 2016). „Synthetic Treebanking for Cross-lingual Dependency Parsing." In: *J. Artif. Int. Res.* 55.1, pp. 209–248.

Tiedemann, Jörg (2012). „Parallel Data, Tools and Interfaces in OPUS." In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur

Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer (2003). „Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network." In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. NAACL '03. Edmonton, Canada, pp. 173–180.

Traffis, Catherine (2019). *Appositives-What They Are and How to Use Them.*

Tsvetkov, Yulia, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer (2016). „Polyglot Neural Language Models: A Case Study in Cross-Lingual Phonetic Representation Learning." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, pp. 1357–1366.

Tyers, Francis M. and Jonathan N. Washington (2015). „Towards a Free/Open-source Universal-dependency Treebank for Kazakh." In: *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pp. 276–289.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). „Attention is All you Need." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5998–6008.

Veldhoen, Sara, Dieuwke Hupkes, and Willem H. Zuidema (2016). „Diagnostic Classifiers Revealing how Neural Networks Process Hierarchical Structure." In: *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016.*

Vrandečić, Denny and Markus Krötzsch (2014). „Wikidata: a free collaborative knowledgebase." In: *Communications of the ACM* 57.10, pp. 78–85.

Vulić, Ivan and Anna Korhonen (Aug. 2016). „On the Role of Seed Lexicons in Learning Bilingual Word Embeddings." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 247–257.

Vulić, Ivan, Nikola Mrkšić, and Anna Korhonen (2017). „Cross-Lingual Induction and Transfer of Verb Classes Based on Word Vector Space Specialisation." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Wada, Takashi, Tomoharu Iwata, and Yuji Matsumoto (July 2019). „Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 3113–3124.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2018). „GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." In: *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pp. 353–355.

Wang, Dingquan and Jason Eisner (2018). „Synthetic Data Made to Order: The Case of Parsing." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 1325–1337.

Wang, Weiyue, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney (2016). „CharacTer: Translation Edit Rate on Character Level." In: *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pp. 505–510.

Wei, Xiang, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang (2018). „Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect." In: *Proceedings of ICLR*.

Wu, Shijie and Mark Dredze (Nov. 2019). „Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 833–844.

Xing, Chao, Dong Wang, Chao Liu, and Yiye Lin (2015). „Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation." In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado, pp. 1006–1011.

Xu, Ruochen, Yiming Yang, Naoki Otani, and Yuexin Wu (2018). „Unsupervised Cross-lingual Transfer of Word Embedding Spaces." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 2465–2474.

Yamada, Ikuya, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji (2017). „Learning Distributed Representations of Texts and Entities from Knowledge Base." In: *Transactions of the Association for Computational Linguistics* 5, pp. 397–411.

Yang, Jiaolong, Hongdong Li, Dylan Campbell, and Yunde Jia (2016). „Go-ICP: A Globally Optimal Solution to 3D ICP Point-Set Registration." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.11.

Yang, Pengcheng, Fuli Luo, Shuangzhi Wu, Jingjing Xu, and Dongdong Zhang (2019). „Learning Unsupervised Word Mapping via

Maximum Mean Discrepancy.“ In: *Natural Language Processing and Chinese Computing*. Ed. by Jie Tang, Min-Yen Kan, Dongyan Zhao, Sujian Li, and Hongying Zan. Cham: Springer International Publishing, pp. 290–302.

Youn, Hyejin, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya (2016). „On the universal structure of human lexical semantics.“ In: *Proceedings of the National Academy of Sciences* 113.7, pp. 1766–1771.

Zampieri, Marcos, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli (2017). „Findings of the VarDial Evaluation Campaign 2017.“ In: *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Valencia, Spain, pp. 1–15.

Zellers, Rowan, Yonatan Bisk, Ali Farhadi, and Yejin Choi (2019). „From Recognition to Cognition: Visual Commonsense Reasoning.“ In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6720–6731.

Zellers, Rowan, Yonatan Bisk, Roy Schwartz, and Yejin Choi (2018). „SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference.“ In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 93–104.

Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov (2018). „CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.“ In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium, pp. 1–21.

Zeman, Daniel and Philip Resnik (2008). „Cross-Language Parser Adaptation between Related Languages.“ In: *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Zeman, Daniel et al. (2017). „CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.“ In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada, pp. 1–19.

Zhang, Meng, Yang Liu, Huanbo Luan, and Maosong Sun (July 2017). „Adversarial Training for Unsupervised Bilingual Lexicon Induction.“ In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada, pp. 1959–1970.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić.

Copenhagen, Denmark 2021