ACL 2019

TyP-NLP: The Workshop on Typology for Polyglot NLP

Proceedings of the First Workshop

August 1, 2019
Florence, Italy


Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-950737-29-1

Introduction

The TyP-NLP workshop will be the first dedicated venue for typology-related research and its integration into multilingual Natural Language Processing (NLP). The workshop will be hosted by the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019) in Florence, Italy.

The ultimate goal of TyP-NLP is the development of robust language technology applicable across the world's languages. Long overdue, the workshop is specifically aimed at raising awareness of linguistic typology and its potential in supporting and widening the global reach of multilingual NLP. It will foster research and discussion on open problems relevant to the multilingual NLP community, but it will also invite input from leading researchers in linguistics and cognitive sciences.

The final program of TyP-NLP contains 5 keynote talks and 18 accepted posters, selected among a large number of non-archival submissions. This workshop would not have been possible without the hard work of its program committee. We would like to express our gratitude to them for writing meticulous reviews in a very constrained span of time. We should also thank our invited speakers, Isabelle Augenstein, Emily Bender, Balthasar Bickel, Jason Eisner, and Sabine Stoll, for their irreplaceable contribution to our program. The workshop is generously sponsored by Google and by the European Research Council (ERC) Consolidator Grant LEXICAL (no. 648909).

Find more details on the TyP-NLP 2019 website: `https://typology-and-nlp.github.io/`

Organizing Committee:

Haim Dubossarsky, University of Cambridge
Arya D. McCarthy, Johns Hopkins University
Edoardo M. Ponti, University of Cambridge
Ivan Vulić, University of Cambridge
Ekaterina Vylomova, University of Melbourne

Steering Committee:

Yevgeni Berzak, MIT
Ryan Cotterell, University of Cambridge
Manaal Faruqui, Google
Eitan Grossman, Hebrew University of Jerusalem
Anna Korhonen, University of Cambridge
Thierry Poibeau, Centre national de la recherche scientifique (CNRS)
Roi Reichart, Technion - Israel Institute of Technology

Program Committee:

Željko Agić, University of Copenhagen
Ehsaneddin Asgari, University of California, Berkeley
Barend Beekhuizen, University of Toronto
Christian Bentz, University of Zürich
Richard Futrell, UC Irvine
Edward Gibson, MIT
Giuseppe Celano, Leipzig University
Mans Hulden, University of Colorado
Gerhard Jäger, University of Tuebingen
Katharina Kann, NYU
Douwe Kiela, Facebook
Silvia Luraghi, University of Pavia
John Mansfield, University of Melbourne
David R. Mortensen, Carnegie Mellon University
Phoebe Mulcaire, University of Washington
Jason Naradowsky, Johns Hopkins
Joakim Nivre, Uppsala University
Robert Östling, Stockholm University
Ella Rabinovich, University of Toronto
Michael Regan, University of New Mexico
Tanja Samardzic, University of Zurich
Hinrich Schütze, University of Munich
Sabine Stoll, University of Zurich
Jörg Tiedemann, University of Helsinki
Reut Tsarfaty, Open University of Israel
Yulia Tsvetkov, Carnegie Mellon University

Daan van Esch, Google
Shuly Wintner, University of Haifa

Invited Speakers:

Isabelle Augenstein, University of Copenhagen
Emily M. Bender, University of Washington
Balthasar Bickel, University of Zurich
Jason Eisner, Johns Hopkins University
Sabine Stoll, University of Zurich

Panelists:

Timothy Baldwin, University of Melbourne
Emily M. Bender, University of Washington
Balthasar Bickel, University of Zurich
Kilian Evang, University of Düsseldorf
Jungo Kasai, University of Washington
Yova Kementchedjhieva, University of Copenhagen
Sabine Stoll, University of Zurich

Conference Program

Thursday, August 1, 2019

7:30–8:45 ***Breakfast***

8:45–9:00 ***Opening Remarks***

9:00–9:45 *Invited Talk*
Emily M. Bender

9:45–10:30 *Invited Talk*
Jason Eisner

10:30–11:00 ***Coffee Break***

11:00–12:30 **Poster Session**

Unsupervised Document Classification in Low-resource Languages for Emergency Situations
Nidhi Vyas, Eduard Hovy and Dheeraj Rajagopal

Polyglot Parsing for One Thousand and One Languages (and Then Some)
Ali Basirat, Miryam de Lhoneux, Artur Kulmizev, Murathan Kurfalı, Joakim Nivre and Robert Östling

Dissecting Treebanks to Uncover Typological Trends: a Multilingual Comparative Approach
Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi

What Do Multilingual Neural Machine Translation Models Learn about Typology?
Ryokan Ri and Yoshimasa Tsuruoka

Syntactic Typology from Plain Text Using Language Embeddings
Taiqi He and Kenji Sagae

Typological Feature Prediction with Matrix Completion
Annebeth Buis and Mans Hulden

Feature Comparison across Typological Resources
Tifa de Almeida, Youyun Zhang, Kristen Howell and Emily M. Bender

Using Typological Information in WALS to Improve Grammar Inference
Youyun Zhang, Tifa de Almeida, Kristen Howell and Emily M. Bender

Cross-linguistic Robustness of Infant Word Segmentation Algorithms: Oversegmenting Morphologically Complex Languages
Georgia R. Loukatou

Cross-linguistic Semantic Tagset for Case Relationships
Ritesh Kumar, Bornini Lahiri and Atul Kr. Ojha

AfricaSign - a Crowd-sourcing Platform for Lexical Documentation of African Sign Languages
Abdelhadi Soudi, Kristof Van Laerhoven and Elmosta Bou-Souf

Predicting Continuous Vowel Spaces in the Wilderness
Emily Ahn and David R. Mortensen

Towards a Multi-view Language Representation: A Shared Space of Discrete and Continuous Language Features
Arturo Oncevay, Barry Haddow and Alexandra Birch

Cross-lingual CCG Induction: Learning Categorial Grammars via Parallel Corpora
Kilian Evang

Towards a Computationally Relevant Typology for Polyglot/Multilingual NLP
Ada Wan

Transfer Learning for Cognate Identification in Low-Resource Languages
Eliel Soisalon-Soininen and Mark Granroth-Wilding

Contextualization of Morphological Inflection
Ekaterina Vylomova, Ryan Cotterell, Timothy Baldwin, Trevor Cohn and Jason Eisner

Towards Unsupervised Extraction of Linguistic Typological Features from Language Descriptions
Søren Wichmann and Taraka Rama

12:30–14:00 ***Lunch***

14:00–14:45 *Invited Talk*
Balthasar Bickel

14:45–15:30 *Invited Talk*
Sabine Stoll

Thursday, August 1, 2019 (continued)

15:30–16:00 ***Coffee Break***

16:00–16:45 *Invited Talk*
Isabelle Augenstein

16:45–17:30 ***Panel Discussion***

17:30–17:45 ***Best Paper Announcement and Closing Remarks***

Biography of the Speakers

Isabelle Augenstein is a tenure-track assistant professor at the University of Copenhagen, Department of Computer Science since July 2017, affiliated with the Copenhagen NLP group and the Machine Learning Section, and work in the general areas of Statistical Natural Language Processing and Machine Learning. Her main research interests are weakly supervised and low-resource learning with applications including information extraction, machine reading and fact checking.

Emily M. Bender's primary research interests are in multilingual grammar engineering, the study of variation, both within and across languages, and the relationship between linguistics and computational linguistics. She is the LSA's delegate to the ACL. Her 2013 book *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax* aims to present linguistic concepts in an manner accessible to NLP practitioners.

Jason Eisner works on machine learning, combinatorial algorithms, probabilistic models of linguistic structure, and declarative specification of knowledge and algorithms. His work addresses the question, "How can we appropriately formalize linguistic structure and discover it automatically?"

Balthasar Bickel aims at understanding the diversity of human language with rigorously tested causal models, i.e. at answering the question what's where why in language. What structures are there, and how exactly do they vary? Engaged in both linguistic fieldwork and statistical modeling, he focuses on explaining universally consistent biases in the diachrony of grammar properties, biases that are independent of local historical events.

Sabine Stoll questions how children can cope with the incredible variation exhibited in the approximately 6000–7000 languages spoken around the world. Her main focus is the interplay of innate biological factors (such as the capacity for pattern recognition and imitation) with idiosyncratic and culturally determined factors (such as for instance type and quantity of input). Her approach is radically empirical, based first and foremost on the quantitative analysis of large corpora that record how children learn diverse languages.

Unsupervised Document Classification in Low-resource Languages for Emergency Situations

Nidhi Vyas Eduard Hovy Dheeraj Rajagopal
Carnegie Mellon University
{nkvyas,dheeraj}@andrew.cmu.edu
hovy@cmu.edu

1 Introduction

During emergency, relief workers need to constantly track updates so that they can learn of situations that require immediate attention (Stowe et al., 2016). However, it is challenging to carry out these efforts rapidly when the information is expressed in people's native languages, which have little to no resources for NLP. We aim for building an adaptable language-agnostic system for such emergent situations that can classify incident language documents into a set of relevant fine-grained classes of humanitarian-needs and unrest situations. Our approach requires no language specific feature engineering and rather leverages the semantic difference between generic class features to build a classification framework that supports relief efforts. We assume no knowledge of the incident language, except the commonly available bilingual dictionaries (which tend to be very small or are generated from out-of-domain data such as Bible alignments). First, we obtain keywords for each target class using English news corpora (Naik et al., 2017; Marujo et al., 2015; Wen and Rosé, 2012; Özgür et al., 2005; Tran et al., 2013), that are then translated using the available bilingual dictionary (Zhang et al., 2016; Adams et al., 2017). Second, an unsupervised bootstrapping module enhances the generic keywords by adding incident-specific language-specific keywords (Knopp, 2011; Huang and Riloff, 2013; Ebrahimi et al., 2016). Next, we use all the keywords to generate labeled data. Finally, this data is used to train a downstream document classifier. This entire procedure is language-agnostic because it bypasses the necessity to create training data from scratch. We validate this procedure in a low-resource setup, with 7 distinct languages, showing significant improvements over the baseline by atleast 13 F1 points. To the best of our knowledge, our approach is the first to combine the use of distant supervision from English and in-language semantic bootstrapping for such a low-resource task. We believe our method can be used as a strong benchmark for future developments in low-resource unsupervised classification.

2 Approach

Figure 1 shows the overall architecture of our approach, composed of three primary modules.

Keywords: We use English in-domain corpora viz. Google News and Relief-Web corpus[1] to generate task-specific keywords, such that each keyword is strongly indicative of the underlying class(es) of a document. We cluster the documents based on their classes and use tf-idf to pick *top* 100 candidate words for each class. We then compute a *label affinity score* between each candidate and class labels using cosine-similarity between their corresponding Word2Vec embeddings[2]. In this way, each candidate keyword has a different association strength across all classes, and we only retain the ones above threshold 0.9. The pruned keywords are translated into the incident language using the available bilingual dictionary, dropping the ones that are absent. Finally, we use keyword spotting to label each document with class/es of the keyword/s present in it.

Bootstrapping: Dropping keywords during translation leaves a significant fraction of documents unlabeled. To improve the percentage of labeled documents, we expand the keywords within the incident language using a two-step process. First, we cluster the labeled documents (obtained from keyword spotting) based on their classes. For each word in a cluster, we compute the sum of its tf-idf score across other clusters and its average word

[1] Pre-classified English documents into disaster relief needs and emergency situations (https://reliefweb.int)

[2] https://code.google.com/archive/p/word2vec/

Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, pages 1–3
Florence, Italy, August 1, 2019.

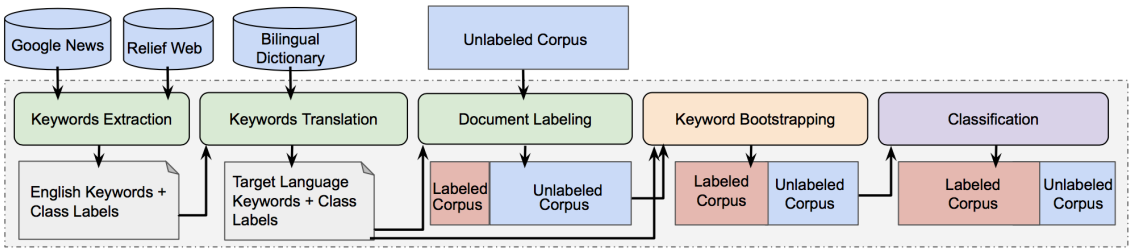

Figure 1: System architecture for low-resource document classification.

	Mandarin			Spanish			Uzbek			Farsi			Tigrinya			Uyghur			Oromo		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Random	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
+ KWD	0.40	0.22	0.28	0.57	0.14	0.23	0.51	0.11	0.18	0.49	0.18	0.26	0.54	0.55	0.55	0.61	0.28	0.39	0.45	0.08	0.14
+ BS	0.34	0.38	0.36	0.22	0.74	0.34	0.33	0.47	0.39	0.31	0.30	0.30	0.53	0.59	0.56	0.52	0.31	0.39	0.33	0.09	0.14
+ KNN	0.31	0.40	0.35	0.31	0.46	**0.37**	0.29	0.47	0.36	0.24	0.48	0.32	0.54	0.58	0.56	0.38	0.36	0.37	0.10	0.12	0.11
SVM	0.30	0.39	0.34	0.30	0.45	0.36	0.34	0.48	**0.40**	0.29	0.29	0.29	0.55	0.58	0.57	0.47	0.38	0.41	0.20	0.09	0.12
R-Forest	0.37	0.38	**0.38**	0.31	0.46	**0.37**	0.33	0.47	0.39	0.23	0.40	0.30	0.56	0.61	0.58	0.42	0.38	0.40	0.12	0.10	0.11
Log-Reg	0.29	0.39	0.33	0.31	0.45	**0.37**	0.32	0.47	0.38	0.23	0.37	0.29	0.56	0.61	0.58	0.51	0.36	**0.42**	0.12	0.10	0.10
GNB	0.32	0.40	0.36	0.30	0.46	0.36	0.33	0.51	**0.40**	0.28	0.31	0.30	0.56	0.62	**0.58**	0.47	0.37	0.41	0.11	0.10	0.11
DAN	0.22	0.38	0.27	0.21	0.74	0.33	0.23	0.52	0.32	0.25	0.68	**0.37**	0.56	0.58	0.57	0.37	0.38	0.37	0.26	0.19	**0.22**
LSTM	0.29	0.39	0.33	0.43	0.23	0.30	0.29	0.48	0.37	0.25	0.24	0.24	0.51	0.63	0.56	0.37	0.31	0.33	0.09	0.20	0.13

Table 1: Results of classification across 7 languages, over each module (Modules - KWD:Keywords, BS:Bootstrap; Classifiers - R-Forest:Random Forest, GNB:Gaussian Naive Bayes, Log-Reg:Logistic regression, LSTM (Hochreiter and Schmidhuber, 1997) and DAN (Iyyer et al., 2015; Chen et al., 2016)

similarity with all other keywords present in that cluster. Each keyword belongs to the same class as its cluster. Second, we prune words with less than 0.9 score, and use the rest to again label more documents using keyword spotting (e.g. fraction of labeled documents in Uzbeck increased by 36%).

Classification: Finally, we use all labelled documents obtained from Keywords and Bootstrapping module to train a classifier, which classifies all the remaining incident language documents.

3 Experiments and Results

We used the LDC corpora[3] for 7 low-resource languages having 11 class labels: *Crime-violence, Terrorism, Regime-change, Medical, Food, Water, Evacuation, Shelter, Search-rescue, Infrastructure,* and *Utilities*. Mandarin, Uzbek, Farsi and Spanish have 190 documents with average 2.7 labels per document. Tigrinya, Uyghur and Oromo have 1.1K, 3.6K and 2.7K documents with 1.4, 0.1 and 1.0 labels per document respectively. Apart from the difference in language families and writing scripts, morphological complexity adds further challenge to classification. As shown in Table 1[4], the *Keywords* module results in highest average F1 gain over the baseline, showing the effectiveness of using language-agnostic information for tasks. As expected, this module primarily improves precision. Further, on average 84.53% keywords were dropped in translation, suggesting improvements in bilingual dictionary can benefit this module. Similarly, the *Bootstrap* module focuses on incident-specific information and primarily improves recall, resulting in an overall F1 gain of 6%. Finally, the *Classifier* module achieves an overall improvement of 4% F1. We observe low performance on Oromo, which is a morphologically rich language. On finer inspection, we found the corpus had several misspelled words. For instance, we identified different versions of *Ethiopia*, such as *itiyoophiyaa*. We also observe that the languages of same family like Uyghur and Uzbek have similar performances. In most cases, the gain in F1 provided by keyword extraction and bootstrapping is significantly higher than that from any classifier. This suggests that the classifier performance will improve only when we improve the mappings between source and target languages.

[3]https://www.ldc.upenn.edu

[4]We use the LOREHLT evaluation guidelines (https://goo.gl/ZT7sMq) for scoring

Acknowledgments

We acknowledge NIST for coordinating the SF type evaluation and providing the test data. NIST serves to coordinate the evaluations in order to support research and to help advance the state-of-the-art. NIST evaluations are not viewed as a competition, and such results reported by NIST are not to be construed, or represented, as endorsements of any participants system, or as official findings on the part of NIST or the U.S. Government. This project was sponsored by the Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O), program: Low Resource Languages for Emergent Incidents (LORELEI), issued by DARPA/I2O under Contract No. HR0011-15-C-0114.

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 937–947.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614*.

Javid Ebrahimi, Dejing Dou, and Daniel Lowd. 2016. Weakly supervised tweet stance classification by relational bootstrapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1012–1017. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ruihong Huang and Ellen Riloff. 2013. Multi-faceted event recognition with bootstrapped dictionaries. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 41–51.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691.

Johannes Knopp. 2011. Extending a multilingual lexical resource by bootstrapping named entity classification using wikipedia's category system. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*, pages 35–43. Asian Federation of Natural Language Processing.

Luis Marujo, Wang Ling, Isabel Trancoso, Chris Dyer, Alan W Black, Anatole Gershman, David Martins de Matos, João Neto, and Jaime Carbonell. 2015. Automatic keyword extraction on twitter. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 637–643. Association for Computational Linguistics.

Aakanksha Naik, Chris Bogart, and Carolyn Rose. 2017. Extracting personal medical events for user timeline construction using minimal supervision. *BioNLP 2017*, pages 356–364.

Arzucan Özgür, Levent Özgür, and Tunga Güngör. 2005. Text categorization with class-based and corpus-based keyword selection. In *International Symposium on Computer and Information Sciences*, pages 606–615. Springer.

Kevin Stowe, Michael J. Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. 2016. Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 1–6. Association for Computational Linguistics.

Dang Tran, Cuong Chu, Son Pham, and Minh Nguyen. 2013. Learning based approaches for vietnamese question classification using keywords extraction from the web. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 740–746.

Miaomiao Wen and Carolyn Penstein Rosé. 2012. Understanding participant behavior trajectories in online health support groups using automatic extraction methods. In *Proceedings of the 17th ACM international conference on Supporting group work*, pages 179–188. ACM.

Dongxu Zhang, Boliang Zhang, Xiaoman Pan, Xiaocheng Feng, Heng Ji, and XU Weiran. 2016. Bitext name tagging for cross-lingual entity annotation projection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 461–470.

Dissecting Treebanks to Uncover Typological Trends. A Multilingual Comparative Approach

Chiara Alzetta[•◇], **Felice Dell'Orletta**[◇], **Simonetta Montemagni**[◇], **Giulia Venturi**[◇]
[•]DIBRIS, Università degli Studi di Genova, Italy
chiara.alzetta@edu.unige.it
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) - ItaliaNLP Lab
{name.surname}@ilc.cnr.it

Introduction and Motivation. Over the last years, linguistic typology started attracting the interest of the community working on cross- and multi-lingual NLP as a way to tackle the bottleneck deriving from the lack of annotated data for many languages. Typological information is mostly acquired from publicly accessible typological databases, manually constructed by linguists. As reported in Ponti et al. (2018), despite the abundant information contained in them for many languages, these resources suffer from two main shortcomings, i.e. their limited coverage and the discrete nature of features (only "the majority value rather than the full range of possible values and their corresponding frequencies" is reported). Corpus-based studies can help to automatically acquire quantitative typological evidence which might be exploited for polyglot NLP. Recently, the availability of corpora annotated following a cross-linguistically consistent annotation scheme such as the one developed in the Universal Dependencies project is prompting new comparative linguistic studies aimed to identify similarities as well as idiosyncrasies among typologically different languages (Nivre, 2015). The line of research described here is aimed at acquiring quantitative typological evidence from UD treebanks through a multilingual contrastive approach.

Method. The proposed methodology is inspired by Alzetta et al. (2018) where an algorithm originally developed for assessing the plausibility of automatically produced syntactic annotations was used to infer quantitative typological evidence from treebanks. The authors demonstrate that the linguistic properties used by this algorithm to rank dependency annotations from reliable to unreliable ones can also be effectively used against manually revised corpora, i.e. gold treebanks. In this case, the resulting ranking of gold dependencies turned out to closely reflect the degree of prototypicality of dependency relations in the target corpus. In this study, we rely on the same algorithm, LISCA (Dell'Orletta et al., 2013), which operates in two steps (see Figure 1): i) it creates a language model collecting statistics about a set of linguistically-motivated features extracted from the automatically parsed sentences of a large reference corpus, and ii) it uses the model to assign a score to each dependency arc contained in a target corpus. Rather than the plausibility of the annotation, the score should be seen here as reflecting the prototypicality degree of a given relation, based on wide variety of features including its context of occurrence. The higher the score of a ranked arc, the more prototypical is the arc with respect to the statistics acquired from the large reference corpus. In Alzetta et al. (2018), the algorithm was used to acquire typological evidence from treebanks relying on LISCA models (LM) of the same language. The main novelty of this study consists in the adopted multilingual comparative approach through which typological evidence is acquired. As illustrated in Figure 2, we ranked the same monolingual treebank using LMs built for different languages, thus obtaining four different dependencies rankings of the same monolingual

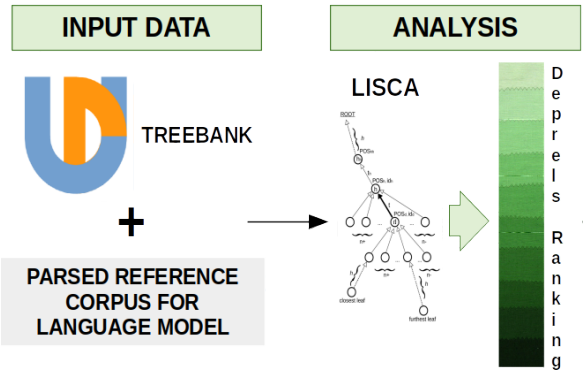

Figure 1: LISCA work-flow.

Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, pages 4–6
Florence, Italy, August 1, 2019.

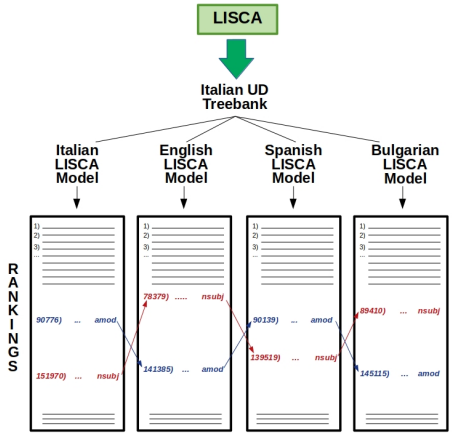

Figure 2: Method work–flow exemplified on IUDT.

LM	Target UDT models			
	IT	EN	SP	BUL
IT	**1.00**	0.94	0.97	0.79
EN	0.84	**1.00**	0.87	0.90
SP	0.98	0.95	**1.00**	0.93
BUL	0.79	0.91	0.83	**1.00**

Table 1: Spearman's correlation between pairs of ranked *amod* using different LMs on each treebank.

LM	Prenominal		Postnominal	
	Up	Down	Up	Down
ENG	21,691.18	2,566.57	290.13	40,934.52
SP	1,541.31	9,163.83	5,462.83	5,322.66
BUL	30,597.76	967.66	885.15	43,003.64

Table 2: Avg difference of ranking positions of IUDT pre- and post-nominal *amod* in different rankings with respect to the ranking obtained with the Italian LM.

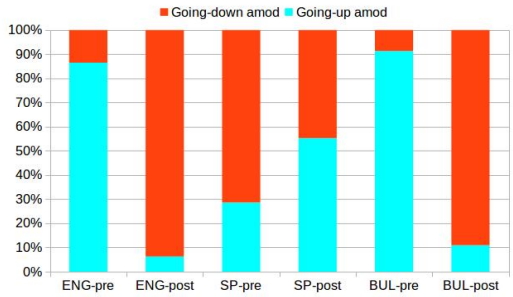

Figure 3: Distributions of the IUDT pre- and post-nominal *amod* fluctuating in different rankings with respect to the ranking obtained with the Italian LM.

treebank. Different positions of the same dependency relation (DR) across the four rankings reflect different degrees of prototypicality of that DR instance: a bigger ranking difference associated with the same DR is connected with stronger typological differences of the languages represented in the selected treebanks, whereas closer rankings reflect typological closeness of languages.

Data. In this study, we considered four UD treebanks (v2.2) (Nivre et al., 2017): English (Silveira et al., 2014), Italian (Bosco et al., 2013), Spanish (McDonald et al., 2013) and Bulgarian (Simov et al., 2005). Statistics to build the LMs for the examined languages were extracted from four monolingual corpora of around 40 million tokens each parsed by the UDPipe pipeline (Straka et al., 2016) trained on the UD treebanks.

Results. Due to space constraints, the methodology is illustrated here wrt Italian UD Treebank (IUDT) and in particular wrt an individual DR: adjectival modifier (*amod*). Table 1 reports Spearman's rank correlation coefficients (p <0.00) obtained through pairwise comparisons of *amod* DRs across the LISCA rankings. Each pair is represented by the ranking obtained using the LM of the target UD treebank (*Target UDT models*) and the ranking obtained using one of the other LMs. Interestingly, typologically similar languages such as IT and SP show higher correlation values (0.98 and 0.97 respectively) than typologically distant ones (e.g. BUL and IT). A similar trend is observed considering the dependency direction of *amod* and its syntactic head. Table 2 and Figure 3 report *i)* the average difference of positions of pre- and post-nominal adjectival modifiers in the IUDT ranking obtained using the Italian LM and the other LMs, and *ii)* the percentage distribution of fluctuations across rankings. It results that higher the number of ranking fluctuations, the more typologically distant the languages are. Namely, in the rankings obtained using LMs of EN and BUL, a higher percentage of prenominal *amod* goes up with respect to the ranking obtained using IT LM. This reflects the linguistic properties used to build LM: right-headed adjectives are more prototypical in EN and BUL than in IT, accordingly they are highly scored by LISCA. As a consequence, a higher percentage of pre-nominal *amod* goes up in the rankings obtained using the EN (86.50) and BUL (91.35) LMs wrt the ranking obtained using IT LM. In addition, the average difference of positions of prenominal adjectives going up in the rankings obtained with EN and BUL LMs is higher (21,691.18 and 30,597.76), as well as the difference of the going-down ranking fluctuations (40,934.52 and 43,003.64). This latter result reflects the lower degree of prototipicality of left-headed adjectives in EN and BUL wrt IT.

References

C. Alzetta, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2018. Universal dependencies and quantitative typological trends. a case study on word order. In *Proceedings of the 11th Edition of International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4540–4549. Association for Computational Linguistics.

C. Bosco, S. Montemagni, and M. Simi. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria.

F. Dell'Orletta, G. Venturi, and S. Montemagni. 2013. Linguistically-driven selection of correct arcs for dependency parsing. *Computaciòn y Sistemas*, 2:125–136.

R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. Bertomeu Castelló, and J. Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

J. Nivre. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing - Proceedings of the 16th International Conference, CICLing 2015, Part I*, pages 3–16, Cairo, Egypt.

J. Nivre, A. Željko, A. Lars, and et alii. 2017. Universal dependencies 2.0. In *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University*.

E.M. Ponti, H. O'Horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, and A. Korhonen. 2018. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *arXiv preprint arXiv:1807.00914*.

N. Silveira, T. Dozat, M.C. de Marneffe, S. Bowman, M. Connor, J. Bauer, and C.D. Manning. 2014. A gold standard dependency corpus for english. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

K. Simov, P. Osenova, A. Simov, and M. Kouylekov. 2005. Design and Implementation of the Bulgarian HPSG-based Treebank. *Journal of Research on Language and Computation. Special Issue*, pages 495–522.

M. Straka, J. Hajic, and J. Strakova. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

What do Multilingual Neural Machine Translation Models learn about Typology?

Ryokan Ri and Yoshimasa Tsuruoka
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{li0123,tsuruoka}@logos.t.u-tokyo.ac.jp

1 Introduction

Unlike traditional statistical machine translation, neural machine translation (NMT) has enabled translation between multiple languages using a single model (Ha et al., 2016; Johnson et al., 2017). It enjoys easier maintainability and production deployment, without changing the model architecture and hurting performance so much.

However, its simplicity raises a natural question: how the multi-lingual model handle the multilingualism? Recent work has shown that the representations learned in NMT models encode a great deal of linguistic information, such as morphology (Belinkov et al., 2017a; Bisazza and Tump, 2018), syntax (Shi et al., 2016), and semantics (Belinkov et al., 2017b; Poliak et al., 2018). However, most analyses are for models which translate in one direction, and only little is known about the linguistic competence of multilingual NMT models.

The goal of this work is to understand the following question: how much do multilingual NMT models capture the universality and variation of languages? Specifically, we will answer the following questions in this paper:

- How much typological information does each module in a model contain?
- How do the architectural choice of a model, specifically the subword or character model, affect the ability to capture linguistic typology?

The experimental design in this work is similar to Malaviya et al. (2017), where they trained a many-to-one multilingual NMT system to predict the typological features of the languages. However, we differ in that, while they focus on learning representation that can be used to predict missing features in typological databases, we aim to analyze the linguistic property of multilingual NMT models and conducted more fine-grained analyses.

2 Methodology

2.1 NMT Training

For a fair comparison among languages, we need sentences aligned in multiple languages. In this experiment, we use the Bible Corpus (Christodouloupoulos and Steedman, 2015), which contains translations of the Bible in 100 languages. We extracted verses aligned among 58 languages, and then split them into train/dev/test sets, which resulted in 23,555/455/455 verses respectively. The test data is used afterward for the following typological prediction task.

The NMT model is the attentional encoder-decoder model similar to Luong et al. (2015). The model has two stacked LSTM layers for the encoder and decoder, and the sizes of embeddings and hidden states are set to 500. The model is trained in the many-to-one scheme, *i.e.*, translating from multiple languages into one single target language. Following Johnson et al. (2017), we do not explicitly specify the source language which the model is translating.

Sentences are segmented by sentencepiece (Kudo, 2018), a language-agnostic tokenizer. For the source languages (57 languages in total), we created a shared vocabulary with the size of 32,000. For the target language (English), the vocabulary size is set to 8,000. We also experimented with character tokenization.

2.2 Probing Task

We investigate the extent to which the NMT model captures the typology of the source languages. We use the URIEL Typological Database (Littell et al., 2017), which compiles typological features of languages extracted from multiple linguistics sources. We used the data

Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, pages 7–9
Florence, Italy, August 1, 2019.

where missing values are predicted with kNN regressions based on phylogenetic or geographical neighbors. We only use the syntactical features, which amount to 103 features, from the database (*e.g.*, S_SUBJECT_BEFORE_VERB, S_PLURAL_PREFIX), as we focus on features that would be directly learned in translation.

Our approach utilizes a probing task (Adi et al., 2017; Conneau et al., 2018). We trained binary logistic regression classifiers dedicated for each typological feature. The classifiers are asked to predict the typological feature of the source languages based on a sentence representation extracted from the trained model (max-pooling of hidden states). We performed 10-fold cross-validation, with no overlapping of languages in each train/test set. As the data in the test set is of the languages unknown to the classifier, the accuracy indicates how much the extracted representation generalizes about the typological features across languages.

3 Results and Discussion

This section presents the results for the two questions we asked: How much typological information does each module in a model contain?; How do the different architectures of a model, specifically subwords or characters model, affect the ability to capture linguistic typology? Table 1 summarizes the result with the majority baseline, where classifiers always predict the majority class for the typological feature.

	Encoder	*Decoder*	*Attention*
majority	80.90%		
subword	84.90%	80.10%	81.30%
character	87.00%	80.10%	84.90%

Table 1: The accuracy of typology prediction using features extracted from different layers of the model. The values are averaged across all the predicted typological features and languages. *Encoder* and *Decoder* represents the output from the top layer of the encoder and decoder respectively. The lower layers gave lower accuracy in most cases. *Attention* is the representation after computing attention, before the output projection layer.

Effect of module

The representations from the encoder predict the typological feature of the source language significantly better than the majority baseline, whereas the decoder sees almost no improvements from the baseline. This indicates the encoder is aware of what language it encodes, whereas the decoder is ignorant of the source and focuses on generating the target language.

However, although the decoder is unaware of the source language, the representation from the attention, again, contains the typological information on the source language. This indicates the inefficiency of the current shared-attention architecture. Ideally, to achieve the most efficient parameter sharing in multilingual translation systems, target sequence generation should be ignorant of the source language properties, as the sentences with the same meaning are eventually mapped to the same target sequence regardless of the source language. In other words, the decoder has to generate a word sequence based on *interlingua*, *i.e.*, shared meaning representation across all languages (Richens, 1958; Schwenk and Douze, 2017; Johnson et al., 2017). Our result confirms that attention is one of the obstacles to language-agnostic generation of the decoder, and is in line with recent efforts to improve multilingual NMT by seeking *neural interlingua* (Lu et al., 2018; Cífka and Bojar, 2018).

Subword vs. characters

The character model is more predictive of typological properties than the subword model in every layer. This can be attributed to the ability of the character model to capture morphology (Qian et al., 2016; Belinkov et al., 2017a), which is verified by the top 5 typological features that see improvement from the subword model to the character model. Three of them are features concerning part-of-speech (ADJECTIVE and NOUN), and one is about dependency marking:

- S_ADJECTIVE_AFTER_NOUN
- S_ADJECTIVE_BEFORE_NOUN
- S_INDEFINITE_WORD
- S_ADJECTIVE_WITHOUT_NOUN
- S_TEND_DEPMARK.

4 Conclusion

We will continue to experiment with another dataset, larger models, and different target languages to verify the observation in this paper and conduct further analyses. We also intend to use other probing tasks, such as universal part-of-speech tagging and natural language inference, to investigate the generalization ability of multilingual NMT models across languages.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Predicition Tasks. In *Proceedings of the International Conference on Learning Representations*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks . In *Proceedings of the International Joint Conference on Natural Language Processing*.

Arianna Bisazza and Clara Tump. 2018. The Lazy Encoder: A Fine-Grained Analysis of the Role of Morphology in Neural Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Ondřej Cífka and Ondřej Bojar. 2018. Are BLEU and Meaning Representation in Opposition? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. *arXiv.org*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viegas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates . In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Patrick Littell, David Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Conference on Machine Translation*.

Minh-Yhang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning Language Representations for Typology Prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018. On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Investigating Language Universal and Specific Properties in Word Embeddings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

R H Richens. 1958. Interlingual Machine Translation. *The Computer Journal*, 1(3):144–147.

Holger Schwenk and Matthijs Douze. 2017. Learning Joint Multilingual Sentence Representations with Neural Machine Translation. In *Proceedings of the Workshop on Representation Learning for NLP*.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does String-Based Neural MT Learn Source Syntax? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Syntactic Typology from Plain Text Using Language Embeddings

Taiqi He
University of California, Davis
tqhe@ucdavis.edu

Kenji Sagae
University of California, Davis
sagae@ucdavis.edu

1 Introduction

We examine the question of whether information about linguistic typology can be derived automatically solely from text corpora, without access to any kind of annotation or parallel data. We describe an ongoing effort to use text from various languages to develop an unsupervised approach to characterize languages through *language embeddings*, which encode information about the structure of languages as vectors. We then explore whether these language vector representations encode typological information, which would traditionally require human expertise.

In recent work, Malaviya et al. (2017) showed that it is possible to extract language representations from neural machine translation models, while Wang and Eisner (2017) used POS tag sequences to predict dependency orders for various languages. In contrast, we do not leverage any kind of linguistic annotation or parallel data. Bjerva and Augenstein (2018) used an architecture similar to ours to generate language representations and showed improvement in task related WALS predictions after task specific transfer learning. We show that it is possible to obtain results better than baseline without optimizing for specific tasks.

2 Method

Our approach is based on the idea behind a denoising autoencoder (Vincent et al., 2008) applied to many languages simultaneously. Given a text corpus with unrelated sentences in each language, we use an encoder-decoder model that learns to reorder, or denoise, sentences in each language.

We first map the words from the various languages into a common representation, leveraging the multilingual 300-dimensional word embeddings from Facebook project MUSE (Conneau et al., 2017). We then build multilingual dictionaries using CSLS (Conneau et al., 2017) and replace each word in each sentence to its English translation. We therefore have a corpus with sentences consisting of English words in the original orders from the different languages. The words in each sentence of this corpus are reordered randomly, creating the input, or source, sequences. The target sequences are the corresponding original sentences. The model then must learn to reorder words in each language, from a random order, to the original order. Additionally, we provide the model with information about what language the sentence is from by appending to each word on both the source and target sides a feature that corresponds to the language identity. Table 1 shows how our target sentences are represented, with English words, original word orders, and language features, along with the original sentences for comparison. A 50-dimensional embedding of this language identity feature is learned along with the reordering task. The intuition is that the model will learn that reordering the same words is done differently depending on whether the language is English, French, Turkish, Vietnamese, etc., but that certain languages are more similar to, or more different from, each other. Once the model is trained, the language feature embedding that helps the model learn how to reorder words for a specific language is the language embedding. After training, we retrieve the language embeddings from the decoder of the model and examine them in the following section.

Our BiLSTM encoder and LSTM decoder both have two layers of 500 units. We used 29 languages, with 200,000 sentences each. Although the MUSE embeddings used in our experiments were created using bilingual dictionaries, violating our goal of deriving the language representations from text only, we also plan to examine the use of

Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, pages 10–12
Florence, Italy, August 1, 2019.

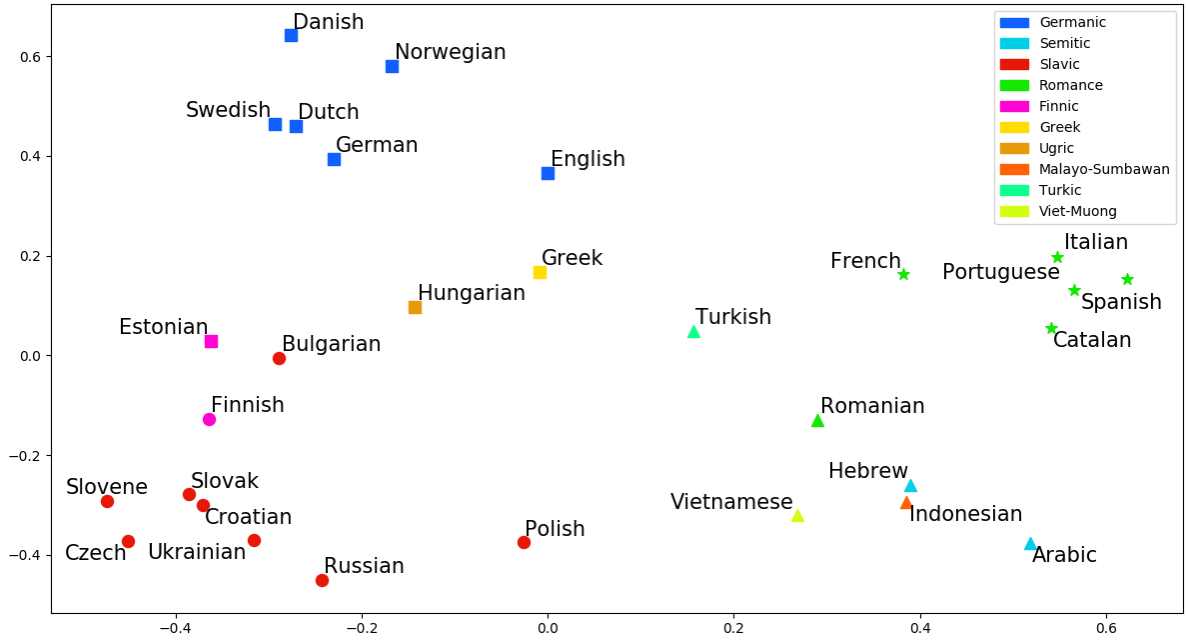

Figure 1: PCA projection of the language embeddings. Shapes represent automatically derived clusters. The Rand score between actual (color) and predicted (shape) language categorizations is .62.

Original	**Transformed**
Er hat den roten Hund nicht gesehen	he|de has|de the|de red|de dog|de not|de seen|de
No vio al perro rojo	not|es saw|es the|es dog|es red|es
Il n'a pas vu le chien rouge	he|fr not|fr has|fr seen|fr the|fr dog|fr red|fr

Table 1: The sentence *He didn't see the red dog* transformed from German, Spanish, and French to English. Word orders were preserved and a label denoting the origin language was attached to each word.

multilingual embeddings obtained from text alone (Lample et al., 2017).

3 Examining Language Embeddings

Figure 1 shows the two-dimensional PCA projection of the normalized language embeddings. We can clearly see clustering of Slavic languages on the lower left, Romance on the right, and Germanic on the upper left. Our dataset also had two Finnic languages, which appear in the upper region in the Slavic languages, and two Semitic languages, which appear on the lower right. The other languages are from families underrepresented in our dataset, and appear either around the center (in the case of Hungarian, Turkish and Greek), or in the lower right corner (Vietnamese, Indonesian). Romanian, a Romance language, appears miscategorized by our language embeddings, also in the lower right cluster.

In addition to actual language relationships, represented by color, we also present the result of spectral clustering with four categories through different shapes. This indicates that, broadly, the language embeddings do capture similarities within language families and dissimilarity across language families. Finally, we train linear models to predict WALS (Dryer and Haspelmath, 2013) features for each language based on the language's 50-dimensional embedding. Even with only 28 training samples (the models were evaluated by leaving one language out), the models predict features under the areas of verbal categories, word order, nominal categories, simple clauses, phonology and lexicon above the level of a majority baseline, with average accuracy of 0.77 (baseline .71), while morphology, nominal syntax and other were worse than or equal to a majority baseline, with average accuracy 0.73 (baseline .76). Despite the surprising failure to capture nominal syntax, it does appear that the language embeddings capture some aspects of syntactic typology.

References

Johannes Bjerva and Isabelle Augenstein. 2018. From Phonology to Syntax: Unsupervised Linguistic Typology at Different Levels with Language Embeddings. *arXiv:1802.09375 [cs]*. ArXiv: 1802.09375.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. *arXiv:1710.04087 [cs]*. ArXiv: 1710.04087.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised Machine Translation Using Monolingual Corpora Only. *arXiv:1711.00043 [cs]*. ArXiv: 1711.00043.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning Language Representations for Typology Prediction. *arXiv:1707.09569 [cs]*. ArXiv: 1707.09569.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.

Dingquan Wang and Jason Eisner. 2017. Fine-Grained Prediction of Syntactic Typology: Discovering Latent Structure with *Supervised* Learning. *Transactions of the Association for Computational Linguistics*, 5:147–161.

Typological Feature Prediction with Matrix Completion

Annebeth Buis
Department of Linguistics
University of Colorado Boulder
anne.buis@colorado.edu

Mans Hulden
Department of Linguistics
University of Colorado Boulder
mans.hulden@colorado.edu

1 Introduction

Currently, there are approximately 7,000 non-extinct languages in the world (Lewis, 2009). Linguistic typology aims at studying and classifying these languages in a systematic way, based on their structural and functional features. The World Atlas of Language Structures (WALS, Dryer and Haspelmath, 2013) is an online database that describes typological features—phonological, syntactic, lexical, word order features, etc.—and records the value of 144 such features for 2,679 languages. However, even for this small set of languages and features, the value of 80% of the language-feature combinations is undefined. Previous research has shown that WALS feature values can be predicted based on the existing data in WALS (e.g., Takamura et al., 2016). Predicted values for missing language-feature combinations in WALS can be useful both for downstream NLP tasks (e.g., Naseem et al., 2012; Daiber et al., 2016) and for typological research.

We discuss predicting WALS data using matrix completion. In our first experiment, we use a simple set-up and a leave-one-out-cross-validation to predict feature values in the database. We compare the results to a majority class baseline and a logistic regression classifier. In further experiments, we test the robustness of our method by leaving out (1) features in the same domain and (2) languages within the same language family. To our knowledge, matrix completion approaches have not been used for typological prediction tasks previously; we show that they outperform our two baselines on the WALS data set.

2 Related Work

For space reasons, we point the reader to Ponti et al. (2018) for a comprehensive overview of research on typological information in NLP/computational linguistics. We have included a comparison to accuracies obtained in other WALS prediction experiments in Table 1. Note that it is difficult to objectively compare performance between different projects because of the wide disparity in methods and subsets of the WALS data used. Because of this, the focus of our comparison will be the two baselines evaluated on the same data set used for matrix completion.

3 WALS and Preprocessing

Features in WALS are split up in 11 domains.[1,2] The data set also includes 10 meta-features (isocodes, language family, genus, etc.), which are not included in the data for prediction. Including meta-features improves accuracy and would therefore be desirable in combination with a downstream NLP task. However, the current set-up is more interesting from a linguistic perspective, since it constrains feature predictions to depend on typological implications (such as *VSO order* → *Noun-Adjective order,* Greenberg, 1963) and other statistical patterns in the WALS data.

The original WALS matrix contains categorical feature values, which were binarized before running matrix completion. We excluded 214 languages for which only 1 feature value has been recorded in WALS. Our method requires no additional preprocessing.[3]

4 Matrix completion

Matrix completion algorithms are not yet part of the standard computational linguistics toolkit. However, there are several reasons why matrix completion is potentially a good method for our

[1] Phonology, sign languages, morphology, nominal categories, nominal syntax, verbal categories, word order, simple clauses, complex sentences, lexicon and other.

[2] Both sign languages and features related to sign languages have been excluded from the data in this project.

[3] All data and code used to obtain the results in this paper is available at https://github.com/annebeth/wals-matrix-completion.

Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, pages 13–15
Florence, Italy, August 1, 2019.

	Accuracy	Method
Georgi et al. (2010)	65.5%	Language clustering
Takamura et al. (2016)	75.5%	Logistic regression
without language family	73.0%	Logistic regression
Murawaki (2017)	74.5%	Bayesian model
Baseline 1	53.1%	Majority class
Baseline 2	65.7%	Logistic regression
Matrix completion	74.3%	IterativeSVD
without domain	61.6%	IterativeSVD
without language family	71.2%	IterativeSVD

Table 1: Matrix completion experiment results compared with results obtained in previous work.

task. First, these algorithms have been used extensively with sparse matrices. Second, since they learn from the entire matrix at once, we expect them to be able to learn more holistic patterns in the data than individual local predictors (such as our logistic regression baseline).

The matrix completion algorithm used in this paper is IterativeSVD,[4] based on Troyanskaya et al. (2001). This method attempts to learn a low-rank approximation of the original matrix by using Singular Value Decomposition (SVD).

5 Experimental set-up

In our experiment, we are predicting each language × feature-combination that currently has a value in WALS (i.e., it is not *undefined*) separately by using leave-one-out cross validation (LOOCV).

First, we calculate results for two baselines. The first baseline predicts a feature value by simply assigning the majority class for each feature. The second baseline consists of logistic regression classifiers that are trained to predict a specific feature based on all other features. Besides our basic matrix completion setting, we have run the experiment in two additional settings: (1) *without domain*-setting: all features from the same domain as the feature that is being predicted are excluded from the matrix, and (2) *without family*-setting: when predicting a feature value for a certain language, all other languages that are in the same language family are excluded from the matrix.

6 Results

Table 1 shows the prediction results obtained with matrix completion on the WALS data and compares them to results obtained in related work.[5] Matrix completion significantly outperforms our two baselines and also improves on the baselines in the *without language family* setting.

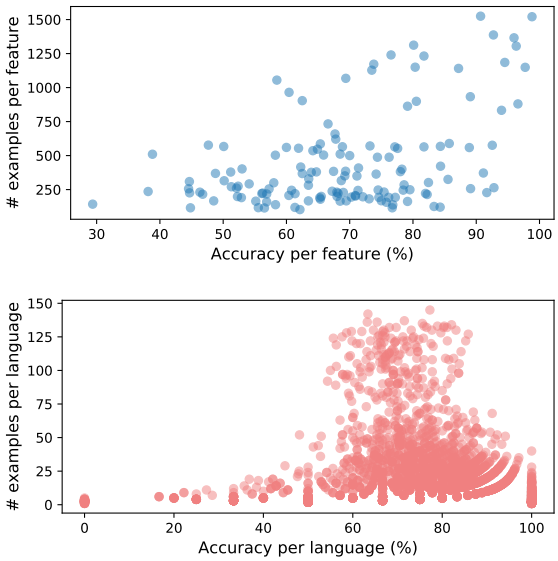

Figure 1: Comparison of (top) number of examples of each feature to the prediction accuracy for that feature and (bottom) the number of examples of each language to the prediction accuracy for that language.

Figure 1 shows different distributional patterns in the comparison of the number of examples with the obtained accuracy. The number of examples shows no correlation with the prediction accuracy per language. For feature accuracy, however, having more examples of a feature can result in better predictions.

7 Conclusion

Matrix completion outperforms the baselines on the WALS data and performs on par with previous work. This shows that matrix completion captures holistic patterns in the data that cannot be learned in a traditional classifier approach. Furthermore, our method requires minimal preprocessing and can easily be used with any typological database.

Our work has shown that treating WALS as a matrix is an effective approach. Daumé III and Campbell (2007) showed that typological implications can be learned from WALS. Non-negative matrix factorization (Lee and Seung, 1999) could be used for the analysis of these implications or to improve the clustering of languages.

[4] IterativeSVD as implemented in the FancyImpute Python package: https://github.com/iskandr/.

[5] We included papers that use only WALS as training data and evaluate on all domains in WALS.

References

Joachim Daiber, Miloš Stanojević, and Khalil Sima'an. 2016. Universal reordering via linguistic typology. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3167–3176.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72.

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

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 385–393.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. Cambridge: MIT Press.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*, sixteenth edition. SIL International, Dallas, TX, USA.

Yugo Murawaki. 2017. Diachrony-aware induction of binary latent representations from typological features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 451–461, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 629–637. Association for Computational Linguistics.

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

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 69–76.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.

Feature Comparison across Typological Resources

Tifa de Almeida, **Youyun Zhang**, **Kristen Howell**, and **Emily M. Bender**

Department of Linguistics, University of Washington, Seattle, WA, USA

{trda, youyunzh, kphowell, ebender}@uw.edu

1 Introduction

We explore in this abstract the relationship between a typological database (WALS) and a grammar specification system (the LinGO Grammar Matrix).

The LinGO Grammar Matrix (GM) customization system automatically generates a custom HPSG grammar based on user inputs to a web-based questionnaire (Bender et al., 2002, 2010). One of its goals is to make the process of creating new grammars easier for linguists of all backgrounds working on testing linguistic hypotheses.

The World Atlas of Linguistics Structures (WALS), often considered a typological atlas due to its detailed geographical data, distinguishes itself from other typological databases such as Terraling (Terraling) by the quantity and the quality of data it contains. WALS was built upon the work of 55 authors who classified over 2,500 languages according to 192 features,[1] and integrated this information with geographical coordinates for each language. It has been used to discover universal typological implications (Daume III and Campbell, 2007), compare phylogenetic relationships within feature-based language clusters (Georgi et al., 2010) and provide a benchmark for automatic typological feature identification from corpora (Lewis and Xia, 2008).

We explore this database in an effort to assist the creation of custom grammars for the GM user. We develope a method to determine what is the overlap between the GM questionnaire and WALS features and how they correspond to each other. In the following sections, we detail how WALS features can be mapped to the Grammar Matrix and what conclusions can be drawn from these mappings. To illustrate this process we provide examples and conclude with a discussion of the potential impacts of these matches and how they may be applied in future work.

2 Methodology

There is a fundamental difference between the features defined in the WALS database and the ones elicited by the GM user interface: While both are built with features extracted from detailed grammars and current typological literature, WALS is fundamentally a reference database of language typology. Accordingly, WALS features classify typological information about a language but are generally not concerned with all the detail that would be required to implement language-specific grammars (e.g. the particular form affixes take).[2]

Due to this, we developed a simple method to determine which WALS features match which GM features to determine to what extent WALS features can be imported and utilized in the grammar customization process. After studying the documentation for a GM feature, for example Adnominal Possession, we examine WALS' inventory of features looking for key terms in titles that correspond to the GM phenomenon, such as Feature 24A Locus of Marking in Possessive Noun Phrases (Nichols and Bickel, 2013). We assess the values of the feature and organize them with the corresponding questionnaire item. An example of this 1-to-1 pairing can be seen in Table 1.

This is where a careful interpretation of the documentation for each system is necessary. WALS utilizes the term locus to refer to a head-dependent marking relationship (Comrie, 2013), designating the possessed noun as the head noun and the possessor as the dependent. The GM refers to the

[1] This dataset is sparse however: different features are specified for different languages.

[2] WALS also contains significant information about linguistic properties outside the morphological, syntactic and semantic information required by the GM, including phonological and lexical features.

Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, pages 16–18
Florence, Italy, August 1, 2019.

Grammar Matrix	WALS	
Morpheme Placement	**Feature 24A**	**Number of Languages**
On the possessum	Possessor is head-marked	78
On the possessor	Possessor is dependent-marked	98
On both the possessor and the possessum	Possessor is double-marked	22
No possessive morphemes appear	Possessor has no marking	32
—	Other types	6

Table 1: Correlation of GM Adnominal Possession morpheme placement options and WALS Feature 24A Locus of Marking in Possessive Noun Phrases. Data available for 236 languages.

possessor and the possessum (Nielsen and Bender, 2018). Therefore we pair the values "on the possessum" (GM) with "possessor is head-marked" (WALS). This analysis is repeated with all values for each feature.

Another point of consideration this particular feature brings up is that after asking about the position of the possessive morpheme, the GM posits further clarifying questions that depend on which option was chosen. If the morpheme appears "on the possessor", the user is asked if it is an affix, separate word or clitic. When the option "on the possessor" is chosen, the GM offers the user the option to add a feature constraint, such as Case. This kind of information, along with the orthography and distribution of the morphemes, is regrettably not provided by WALS and must remain under the user's purview to add.

Below we offer two more examples of WALS/GM features to illustrate the challenges of feature mapping when a one-to-one correspondence cannot be found or is not particularly useful.

2.1 Case Marking Strategy

The core case marking section of the GM (Drellishak, 2008) corresponds to WALS Feature 98A (Comrie, 2013), entitled *Alignment of Case Marking of Full Noun Phrase*. The latter designates the argument abbreviations according to Dixon (1994) – A (agent of a transitive verb), O (object of a transitive verb) and S (subject of an intransitive verb) – whereas WALS uses P (patient) instead of O (Comrie, 1978).

The GM questionnaire gives the user the option to check which case marking strategy is being used and also what each case is called, e.g. ergative. By extracting the values of feature 98A from WALS, the user is given a head-start at this. For the 190 languages specified for this feature in WALS, the GM user has a readily usable source of information for the first of these steps.

2.2 Number of Cases

WALS feature 49A (Iggesen, 2013), defined for 261 languages, maps how many cases a language contains. One would think that knowing the number of cases would be helpful in building a grammar, but it is actually not. Without also knowing what each case is called, this feature could only notify the user that they must manually add N cases, which is not useful for our purposes.

3 Conclusion and Future Work

Having reviewed 33 WALS features, we estimate that about 20 of them (10.4% of the total) can be usefully imported into the GM system to facilitate the grammar generation process for the user. This corresponds to about 8.5% of the GM's grammar specification options.

Our work identifying which features can be mapped and in what way could support an API that extracts the pertinent information from WALS when the user starts a custom grammar. The first page of the user interface asks the user to input the language ISO code, which is also used in the WALS database. The user would be given a choice of importing this information or not, and should they choose to do so would be shown a notification detailing how many features were found.

Additionally, our mapping of WALS to GM features could support work on automatically answering the GM questionnaire. The AGGREGATION project (e.g. Zamaraeva et al., 2019) presently approaches this problem by inferring GM grammar specifications from collections of interlinear glossed text. Where WALS features map to GM features, and WALS values are available for the language at hand, the WALS values can potentially be used to guide grammar specification inference, as explored in Zhang et al. 2019.

References

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Res. Lang. Comput.*, 8(1):23–72.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *COLING-02: Grammar Engineering and Evaluation*.

Bernard Comrie. 1978. Ergativity. In Winfred P. Lehmann, editor, *Syntactic Typology: Studies in the Phenomenology of Language*, pages 329–394. University of Texas Press, Austin.

Bernard Comrie. 2013. Alignment of case marking of full noun phrases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Hal Daume III and Lyle Campbell. 2007. A bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics.

Robert M. W. Dixon. 1994. *Ergativity*. Cambridge University Press, Cambridge.

Scott Drellishak. 2008. Complex case phenomena in the grammar matrix. In *Proceedings of the 15th International Conference on Head-Driven Phrase Structure Grammar, National Institute of Information and Communications Technology, Keihanna*, pages 67–86, Stanford, CA. CSLI Publications.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 385–393, Beijing, China. Coling 2010 Organizing Committee.

Oliver A. Iggesen. 2013. Number of cases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

William D. Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Johanna Nichols and Balthasar Bickel. 2013. Locus of marking in possessive noun phrases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Elizabeth Nielsen and Emily M. Bender. 2018. Modeling adnominal possession in multilingual grammar engineering. In *Proceedings of the 25th International Conference on Head-Driven Phrase Structure Grammar, University of Tokyo*, pages 140–153, Stanford, CA. CSLI Publications.

Terraling. Available online: http://test.terraling.com. Accessed: 26 April, 2019. [link].

Olga Zamaraeva, Kristen Howell, and Emily M. Bender. 2019. Handling cross-cutting properties in automatic inference of lexical classes: A case study of Chintang. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, volume 1 Papers, pages 28–38, Honolulu, Hawai'i.

Youyun Zhang, Tifa de Almeida, Kristen Howell, and Emily M. Bender. 2019. Using typological information in wals to improve grammar inference. Unpublished paper, submitted to TypNLP.

From phonemes to morphemes: relating linguistic complexity to unsupervised word over-segmentation

Georgia Loukatou
Laboratoire de sciences cognitives et de psycholinguistique, Département d'études cognitives
ENS, EHESS, CNRS, PSL University
georgialoukatou@gmail.com

Abstract

Previous work has documented variation in word segmentation performance across languages, with a trend to yield lower scores for languages with elaborate morphological structure. However, segmenting smaller chunks than words, "oversegmenting", is reasonable from a computational point of view. We predict that oversegmentation would be encountered more often in complex languages. In this work in progress, we use a dataset of 9 languages varying in complexity and focus on cognitively-inspired word segmentation algorithms. Complexity is defined by Compression-based, Type-Token Ratio and Word Length metrics. Preliminary results show that a possible relation between morphological complexity and oversegmentation cannot be predicted exactly by none of these metrics, but may be best approximated by word length.

1 Introduction

The issue of word segmentation is open in the NLP community (e.g., Harris (1955)). Its implementations include processing languages with no orthographic word boundaries, such as Chinese and Japanese. It is also a key problem humans face when acquiring language.

Previous work documented variation in the success rate of segmentation across languages, and a trend to yield lower scores for languages with elaborate morphological structure. This is true for both cognitively inspired (Johnson, 2008; Fourtassi et al., 2013; Loukatou et al., 2018)) and other models (Mochihashi et al., 2009; Zhikov et al., 2013; Chen et al., 2011). Evaluation is conventionally based on orthographic word boundaries.

Do these models manage to learn more linguistic structure, that what is actually described in these accuracy scores? Segmenting smaller meaningful chunks than words is reasonable from a computational point of view: morphologically complex languages often feature multimorphemic, long words, and algorithms might break words up into component morphemes, treating frequent morphemes as words. Finding out morphemes might be useful for later linguistic analysis, especially for languages with rich morphological systems , and such morphemes could be used as cues to further bootstrap segmentation. Thus, a "useful" error in segmentation could be oversegmentation (Gervain and Erra, 2012; Johnson, 2008), the percentage of word tokens returned as two or more subparts in the output.

We thus predict that oversegmentation might be encountered more often in complex languages. To test this, we need data from languages varying in complexity. Since there is no standard way to define complexity, for this study, three metrics are used: first, the Moving Average Type-token Ratio (500-word window) (Kettunen, 2014), and second, two versions of compression-based complexity (Szmrecsanyi, 2016)[1]. The two metrics are normalized (0=least complex, 1=most complex) and their average score is attributed to each language. Third, we look at word length, since, in general, longer words could attract more division.

2 Methods

We use the ACQDIV database (Moran et al., 2016) of typologically diverse languages, with transcriptions of infant-directed and -surrounding speech recordings, from Inuktitut (Allen, 1996), Chintang (Stoll et al., 2015), Turkish (Küntay et al., Unpub-

[1] 1st metric: the size of compressed corpus (gzip) divided by the size of raw corpus. 2nd metric: systematic distortion of morphological regularities, so as to estimate the role of morphological information in the corpus. Each word type is replaced with a randomly chosen number. The size of the distorted compressed corpus is then divided by the size of the originally compressed corpus.

Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, pages 19–21
Florence, Italy, August 1, 2019.

lang	compr.	MATTR	w length	% over	% corr	% total
Inu	1	0.90	8.56	51	22	73
Chi	0.56	0.87	4.39	44	24	68
Tur	0.44	0.86	4.92	39	26	65
Yuc	0.42	0.92	3.80	31	27	58
Rus	0.41	0.91	4.47	46	19	65
Ses	0.31	0.86	4.28	44	25	69
Ind	0.28	0.85	4.11	42	25	67
Jap	0.14	0.87	3.94	37	25	62
Eng	0.02	0.39	3.04	6	51	57

Table 1: Complexity scores for the three metrics are given in the first columns. Percentage of average oversegmented, correct word tokens and their sum are then given per language.

lished), Yucatec (Pfeiler, 2003), Russian (Stoll and Meyer, 2008), Sesotho (Demuth, 1992), Indonesian (Gil and Tadmor, 2007) and Japanese (Miyata and Nisisawa, 2010; Nisisawa and Miyata, 2010). In order to compare with a previously studied language, we included the English Bernstein corpus (MacWhinney, 2000).

Several models have been proposed as plausible strategies used by learners retrieving words from input. We used a set of these strategies (Bernard et al., 2018). Two baselines were Base0, treating each sentence as a word, and Base1, treating each phoneme as a word. DiBS[2] (Daland, 2009) implements the idea that unit sequences often spanning phrase boundaries probably span word breaks. FTP[3] (Saksida et al., 2017) measures transitional probabilities between phonemes and cuts depending on a local threshold (relative, FTPr) or a global threshold (absolute, FTPa). Adaptor Grammar (AG) (Johnson, 2008) assumes that learners create a lexicon of minimal, recombinable units and use it to segment the input. AG implements the Pitman-Yor process. Finally, PUDDLE[4] (Monaghan and Christiansen, 2010) is incremental, and learners insert in a lexicon an utterance that cannot be broken down further, and use its entries to find subparts in subsequent utterances. Before segmentation, spaces between words were removed, leaving the input parsed into phonemes, with utterance boundaries preserved.

3 Results

Statistics regarding corpora and results are presented in Table 1. In general, languages had similar oversegmentation scores, (ranging from 31% to 51% if we exclude English), which did not exactly follow their complexity ranking. Performance difference across languages decreased when considering oversegmented tokens as correctly segmented.

4 Discussion

Word length had the best prediction of oversegmentation compared to other metrics, compression and MATTR. This shows that longer words have more alternative parses, and this could explain oversegmentation results better than other properties inherent to morphologically complex languages. That said, a possible relation between morphological complexity and oversegmentation, could not be *exactly* explained by none of these complexity metrics.

It was also observed that there was no absolute ranking of complexity across languages; on the contrary, it would change according to the feature studied. In general, cross-linguistic differences were small for such a typologically distinct dataset of languages. Further research might shed light on whether this behavior is due to linguistic properties common across languages, or a confound (e.g. corpus size).

Moreover, discovering meaningful units is of particular importance to language acquisition models, such as the ones implemented here. Infant word segmentation algorithms are cognitively plausible only if they are cross-linguistically valid and offer useful insights to learn all linguistic structures. It would also be interesting to compare performance of these models to state-of-the-art NLP algorithms, such as HPYLM (Mochihashi et al., 2009) or ESA (Chen et al., 2011).

A limitation of this study is that the current implementation of WordSeg does not only look at oversegmentation cases resulting in meaningful, morpheme-like sub-parts. A next step would be to focus on reasonable oversegmentation errors, even though not all of these corpora have morpheme annotations.

Measuring reasonable errors such as oversegmentation could shed light on the segmentability of morphologically complex languages and the cross-linguistic applicability of models. Further research might include over-, but also undersegmentation errors, when two or more words in the input returned as a single unit in the output.

[2]Diphone Based Segmentation algorithm

[3]Forward Transitional Probabilities algorithm

[4]Phonotactics from Utterances Determine Distributional Lexical Elements

References

Shanley E. M. Allen. 1996. *Aspects of argument structure acquisition in Inuktitut*. Benjamins, Amsterdam.

Mathieu Bernard, Roland Thiolliere, Amanda Saksida, Georgia Loukatou, Elin Larsen, Mark Johnson, Laia Fibla Reixachs, Emmanuel Dupoux, Robert Daland, Xuan Nga Cao, and Alejandrina Cristia. 2018. Wordseg: Standardizing unsupervised word form segmentation from text. *Behavior research Methods*.

Songjian Chen, Yabo Xu, and Huiyou Chang. 2011. A simple and effective unsupervised word segmentation approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Robert Daland. 2009. *Word segmentation, word recognition, and word learning: A computational model of first language acquisition*. Ph.D. thesis, Northwestern University.

Katherine A. Demuth. 1992. Acquisition of sesotho. In Dan Isaac Slobin, editor, *The crosslinguistic study of language acquisition*, volume 3, pages 557–638. Lawrence Erlbaum Associates, Hillsdale, NJ.

Abdellah Fourtassi, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2013. WhyisEnglishsoeasytosegment. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–10.

Judit Gervain and Ramón Guevara Erra. 2012. The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition*, 125(2):263–287.

David Gil and Uri Tadmor. 2007. The mpi-eva jakarta child language database. a joint project of the department of linguistics, max planck institute for evolutionary anthropology and the center for language and culture studies, atma jaya catholic university.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Mark Johnson. 2008. Unsupervised word segmentation for sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27. Association for Computational Linguistics.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Aylin C. Küntay, Dilara Koçbaş, and Süleyman Sabri Taşçı. Unpublished. Koç university longitudinal language development database on language acquisition of 8 children from 8 to 36 months of age.

Georgia Loukatou, Sabine Stoll, Damian Blasi, and Alejandrina Cristia. 2018. Modeling infant segmentation of two morphologically diverse languages. *TALN*.

Brian MacWhinney. 2000. *The CHILDES project: tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ.

Susanne Miyata and Hiro Yuki Nisisawa. 2010. *MiiPro - Tomito Corpus*. Talkbank, Pittsburgh, PA.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. 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: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics.

Padraic Monaghan and Morten H Christiansen. 2010. Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, 37(3):545–564.

Steven Moran, Robert Schikowski, D Pajović, Cazim Hysi, and Sabine Stoll. 2016. The ACQDIV database: Min (d) ing the ambient language. In *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)*, pages 4423–4429.

Hiro Yuki Nisisawa and Susanne Miyata. 2010. *MiiPro - ArikaM Corpus*. Talkbank, Pittsburgh, PA.

Barbara Pfeiler. 2003. Early acquisition of the verbal complex in yucatec maya. *Development of verb inflection in first language acquisition*, pages 379–399.

Amanda Saksida, Alan Langus, and Marina Nespor. 2017. Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, 20(3):1–11.

Sabine Stoll, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Netra P. Paudyal, Manoj Rai, Novel Kishor Rai, Ichchha P. Rai, Taras Zakharko, Robert Schikowski, and Balthasar Bickel. 2015. Audiovisual corpus on the acquisition of chintang by six children.

Sabine Stoll and Roland Meyer. 2008. Audio-visional longitudinal corpus on the acquisition of russian by 5 children.

Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. *Complexity, isolation, and variation*, 57:71.

Valentin Zhikov, Hiroya Takamura, and Manabu Okumura. 2013. An efficient algorithm for unsupervised word segmentation with branching entropy and mdl. *Information and Media Technologies*, 8(2):514–527.

Predicting Continuous Vowel Spaces in the Wilderness

Emily Ahn and **David Mortensen**
Language Technologies Institute
Carnegie Mellon University
{eahn1,dmortens}@cs.cmu.edu

Abstract

We aim to model the acoustic vowel spaces of 24 diverse languages—a subset taken from the CMU Wilderness corpus of Bible recordings. With this model, we test hypotheses from phonological typology. We also expand upon previous work that used formant measurements taken by field linguists, and we use automatic tools to align and extract vowel segments in large-scale recorded speech. This work in progress is at the stage where data has been carefully processed, just prior to implementation of the model.

1 Introduction

Every language has a system of vowels, whether few or many, and understanding how these systems work crosslingually has been a goal in linguistic phonological typology. We are interested in empirically studying vowel spaces from the CMU Wilderness Multilingual Speech Dataset (Black, 2019), a large database of audio recordings from around 700 languages. Cotterell and Eisner (2018) developed a deep generative model to use acoustic formants to predict vowel spaces across a dataset of 223 languages. We aim to expand upon their findings and build a model from a set of languages that contain rich acoustic data per language and many vowel tokens per vowel type.

From this model, we can analyze our results to answer other questions regarding phonological typology. Dispersion Theory predicts that vowel types will be maximally dispersed within the vowel space, and Focalization Theory predicts that vowel types will preferentially be centered around canonical focii (Schwartz et al., 1997). Both theories make predictions about the distribution of centroids of types within formant space, but neither theory makes an explicit prediction about the dispersion of vowel tokens of a given type. In fact, both of these theories suggest that within-type dispersion should be relatively insensitive to other factors, since they treat vowel space as a system of categories. Other theories of vowel spaces, like those based on Evolutionary Phonology and exemplar theory, predict that the dispersion of tokens of each type should be inversely related to the number of contrasting types within a vowel system, since phonological categories can be seen as competing with one another for phonetic space (Vaux and Samuels, 2015). Reduced vowel inventories, in such a theory, are the result of mergers and merged categories take up more phonetic space than either of the categories prior to merger.

The Wilderness corpus provides a unique opportunity to test whether the number of vowel contrasts in a language's phonological inventory predicts the average dispersion of tokens in each vowel type. Rather than just providing idealized tokens of vowels from many languages or many tokens of vowels from few languages, it provides a massive number of tokens from a very large number of languages. If a relationship between number of types and token dispersion does exist on a large scale, it would be important evidence for evolutionary approaches to vowel space typology. If token dispersion is insensitive to the number of vowel types, support would be lent to the dispersion-focalization model.

These findings would be of interest to computational typologists and have implications for low-resource NLP and speech technologies. In addition to this narrow scientific question, this paper would contribute a replicable methodology for extracting vowels in a subset consisting of 24 languages from the public Wilderness corpus of 700 languages. This methdology could be used to extract vowel tokens from the corpus on a much larger scale.

Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, pages 22–24
Florence, Italy, August 1, 2019.

2 Data

The Wilderness corpus comprises of roughly 700 languages of read speech from the New Testament of the Bible, originally scraped from Bible.is.[1]

2.1 Selection of 24 Languages

For this work, we intersected these languages with the PHOIBLE database of crosslingual phonological inventories (Moran and McCloy, 2019). We chose 24 languages where for each integer between 3 and 10, there are three languages (from distinct regions) whose vowel inventory size is that integer. We also used a criterion of choosing languages with the highest automatic alignment scores, as determined by algorithms provided in the Wilderness data, and this list is given in Table 1.

Language	Country	Vowels	Hours
Cebuano	Philippines	3	22
Kabyle	Algeria	3	8
Tena Quechua	Ecuador	3	19
Yupik	United States	4	22
Maranao	Philippines	4	24
Podoko	Cameroon	4	21
Russian	Russia	5	15
Twampa	Ethiopia	5	31
Urarina	Peru	5	31
Hanga	Ghana	6	14
Paumari	Brazil	6	48
Manado Malay	Indonesia	6	25
Komi	Russia	7	17
Sundanese	Indonesia	7	20
Tigrinya	Ethiopia	7	14
Denya	Cameroon	8	15
Huambisa	Peru	8	28
Maithili	India	8	14
Moru	Sudan	9	23
Nomatsigenga	Peru	9	36
Ossetian	Georgia	9	12
Eastern Oromo	Ethiopia	10	24
Maka	Paraguay	10	29
Tamang	Nepal	10	18

Table 1: The 24 languages chosen for this analysis, from the Wilderness data. They are sorted by number of vowel types as determined by PHOIBLE, and we attempted to balance the languages by region.

2.2 Preprocessing

We first obtained phone-level alignments via the tool provided from Festvox.[2] We then manually mapped the phoneme lists from the data with the IPA from PHOIBLE, since there is noise in the pronunciation model. Given vowel alignments, we extracted means of the first few formants using DeepFormants,[3] a tool for formant estimation. All preprocessing scripts will be made publicly available for future analyses on any language from the Wilderness.

A limitation of the data is that each language recording is spoken by few and undocumented speakers. We also anticipate challenges with regards to formant normalization across speaker gender, given that females tend to have higher and a greater range of formants than male speakers even when controlling for vocal tract.

3 Hypotheses

We hypothesize that vowel cloud size is inversely correlated with the number of vowels in a language's inventory. Cloud size will be measured as the level of dispersion in the probabilistic distribution of the two-dimensional vowel space.

4 Methodology

We plan to implement the deep generative models using determinantal point processes (DPP) from (Cotterell and Eisner, 2018) to analyze the formants in vowel spaces of our subset of the Wilderness data. We may choose a different method as well, in order to capture the variety of vowel tokens since the prior work used one representative vowel token per phoneme. We may use cross-entropy to evaluate our generative models, and will present our findings with regards to our hypotheses.

References

Alan W Black. 2019. CMU wilderness multilingual speech dataset. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975. IEEE.

Ryan Cotterell and Jason Eisner. 2018. A deep generative model of vowel formant typology. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 37–46.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Jean-Luc Schwartz, Louis-Jean Boë, Nathalie Vallée, and Christian Abry. 1997. The dispersion-

[1] http://www.bible.is/

[2] http://festvox.org/

[3] https://github.com/MLSpeech/DeepFormants

focalization theory of vowel systems. *Journal of phonetics*, 25(3):255–286.

Bert Vaux and Bridget Samuels. 2015. Explaining vowel systems: dispersion theory vs natural selection. *The Linguistic Review*, 32(3):573–599.

Transfer Learning for Cognate Identification in Low-Resource Languages

Eliel Soisalon-Soininen
Department of Computer Science
University of Helsinki
eliel.soisalon-soininen@helsinki.fi

Mark Granroth-Wilding
Department of Computer Science
University of Helsinki
mark.granroth-wilding@helsinki.fi

Introduction

In our on-going work, we are addressing the problem of identifying cognates across unannotated vocabularies of any pair of languages. In particular, we assume that the languages of interest are low-resource to the extent that no training data whatsoever, even in closely related languages, are available for the task. Instead, we are investigating the performance of language-independent transfer learning approaches utilising training data from a completely unrelated, higher-resource language family. Our results so far suggest that a Siamese convolutional neural network generalises more effectively across language families than baselines.

Cognate identification is a core task in the *comparative method*, a collection of techniques used in historical linguistics, a field closely tied with linguistic typology (Shields, 2011). Cognate information is also useful for applications such as machine translation (Grönroos et al., 2018). In addition, knowledge of cognates is useful for second-language learning (Beinborn et al., 2014).

Cognate identification

In cognate identification, we are essentially given two string sets $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$. The task is to extract those pairs (x, y) in relation R:

$$R = \{(x, y) \in X \times Y |\ x \text{ is cognate with } y\ \}$$

Each element $x \in X$ and $y \in Y$ is a string over alphabets Σ_x and Σ_y respectively. The alphabet sets do not necessarily overlap.

Table 1 illustrates the difficulty of cognate identification. As can be seen, some cognates are straightforward with a similar form and meaning (e.g. *notte–noche*). On the other hand, there is large variation in the degree of similarity in terms of both form and meaning. However, common

Word A	**Word B**	**Meanings**
it: *notte*	es: *noche*	'night'
fi: *huvittava*	et: *huvitav*	'amusing'; 'interesting'
en: *attend*	fr: *attendre*	'attend'; 'wait'
en: *oath*	sv: *ed*	'oath'
fi: *pöytä*	sv: *bord*	'table'
en: *bite*	fr: *fendre*	'bite'; 'split'

Table 1: Examples of cognates with varying degree of similarity in form and meaning.

to all of these examples is that they exhibit *regular sound correspondences*, i.e. word segments regularly occurring in similar positions and contexts (List, 2013; Kondrak, 2012), such as *oa–e* and *th–d* in English–Swedish cognates. Therefore, cognate identification should rely on the identification of such cross-lingual correspondences (i.e. pairs of single characters or substrings).

Most previous work attempts to design such a string similarity metric that would tend to assign a higher score to cognate than unrelated words. Common approaches include extensions of the traditional Levenshtein distance (Levenshtein, 1966) that either assign weights to pairs of symbols according to their phonetic properties (e.g. List, 2013; Kondrak, 2000), or that learn such weights from example cognates (Ciobanu and Dinu, 2014; Gomes and Pereira Lopes, 2011). McCoy and Frank (2018) use weights based on character embeddings.

In contrast to much of previous work, we make no strict assumptions about the degree of similarity in form or meaning that any two cognates should exhibit. Instead, following Rama (2016) and Jäger (2014), we treat regular correspondences as the main driving factor in the cognate relation and attempt to capture these in a completely data-driven manner. We aim to contribute to this line of research by considering the ability of our models to generalise across language families.

Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, pages 25–27
Florence, Italy, August 1, 2019.

Models and experiments

In our experiments, we have trained our models with an etymological database of Indo-European languages (De Melo, 2014), and tested their performance on combinations of three unannotated vocabularies from Sami languages of the Uralic family. We have experimented with two similarity learning models, a Siamese convolutional neural network (S-CNN) based on Rama (2016) and a support vector machine (SVM) based on Hauer and Kondrak (2011), compared with a Levenshtein-distance (LD) baseline (Levenshtein, 1966). In addition, we have experimented with fine-tuning the S-CNN model in order to quantify the benefit of having small amounts of target-language training data.

The Levenshtein distance between two strings is the minimum number of insertions, deletions, and substitutions needed to transform one string to another. It is straightforward to turn this into a similarity metric. For the SVM, word pairs are encoded into vectors of the following features: Levenshtein distance, number of common bigrams, prefix length, lengths of both words, and the absolute difference between the lengths. The S-CNN is a two-input version of a convolutional neural network, where input words are encoded into matrices of concatenated one-hot vectors representing characters. As shown in Figure 1, the network creates a merged representation of a word pair, to be classified as cognate or unrelated.

Figure 2 shows a precision-recall curve for each model, including both a fine-tuned (with 500 target-language training pairs) and unadapted S-CNN. As expected, the fine-tuned S-CNN outperforms the other models. Interestingly, even the unadapted S-CNN simply relying on Indo-European training data outperforms the SVM and LD. This suggests that the S-CNN is able to more effectively capture such aspects of the cognateness relation that carry over across language families.

Work in progress

We are currently investigating approaches to improve target-family performance with unsupervised methods of domain adaptation. One of our lines of work is to use an adversarial approach to making target-family word pair representations more similar to source-family representations, similarly to the method of Tzeng et al. (2017) intended for domain adaptation of images.

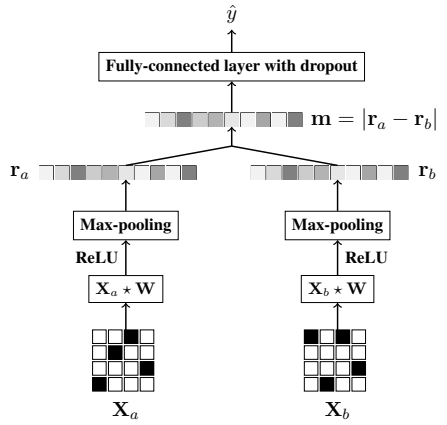

Figure 1: The S-CNN architecture. Column vectors in input matrices represent one-hot-encoded characters. The filter $\mathbf{W}$ is convolved over character sequences.

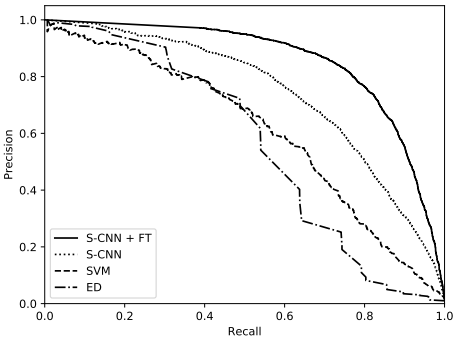

Figure 2: Precision-recall for Sami test set.

Another way to extend the S-CNN model is to use unsupervised multilingual character embeddings (Granroth-Wilding and Toivonen, 2019), trained with small corpora from the target languages. This could be a way to make characters across languages more comparable to each other, thus tackling the issue that orthographies are often not directly comparable.

In addition to unsupervised methods, we also intend to compare our data-driven approaches with more linguistically-informed ones, in order to assess the benefit of such information. For example, our work could benefit from the use of universal phonetic encodings of words instead of orthographic forms.

Although we have thus far specifically focused on the problem of cognate identification, we believe that these methods could be extended to the study of other typological features of language and the automatic inference of such features. Language typology could also provide a means to interpret the representations of the S-CNN model.

References

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Readability for foreign language learning: The importance of cognates. *ITL-International Journal of Applied Linguistics*, 165(2):136–162.

Alina Maria Ciobanu and Liviu P Dinu. 2014. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 99–105.

Gerard De Melo. 2014. Etymological wordnet: Tracing the history of words. In *LREC*, pages 1148–1154. Citeseer.

Luís Gomes and José Pereira Lopes. 2011. Measuring spelling similarity for cognate identification. *Progress in Artificial Intelligence*, pages 624–633.

Mark Granroth-Wilding and Hannu Toivonen. 2019. Unsupervised learning of cross-lingual symbol embeddings without parallel data. *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 19–28.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2018. Cognate-aware morphological segmentation for multilingual neural translation. *arXiv preprint arXiv:1808.10791*.

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of 5th international joint conference on natural language processing*, pages 865–873.

Gerhard Jäger. 2014. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. In *Quantifying Language Dynamics*, pages 155 – 204. Leiden, The Netherlands: Brill.

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 288–295. Association for Computational Linguistics.

Grzegorz Kondrak. 2012. Similarity patterns in words. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 49–53. Association for Computational Linguistics.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Johann-Mattis List. 2013. *Sequence comparison in historical linguistics*. Ph.D. thesis, Heinrich-Heine-Universität Düsseldorf.

Richard T McCoy and Robert Frank. 2018. Phonologically informed edit distance algorithms for word alignment with low-resource languages. *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 102–112.

Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027.

Kenneth Shields. 2011. Linguistic typology and historical linguistics. In *The Oxford Handbook of Linguistic Typology*.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4.

Author Index