A Rich Morphological Tagger for English: Exploring the Cross-Linguistic Tradeoff Between Morphology and Syntax

Christo Kirov[1] **John Sylak-Glassman**[1] **Rebecca Knowles**[1,2] **Ryan Cotterell**[1,2] **Matt Post**[1,2,3]
[1]Center for Language and Speech Processing
[2]Department of Computer Science
[3]Human Language Technology Center of Excellence
Johns Hopkins University
kirov@gmail.com, {jcsg, rknowles, rcotter2}@jhu.edu, post@cs.jhu.edu

Abstract

A traditional claim in linguistics is that all human languages are equally expressive—able to convey the same wide range of meanings. Morphologically rich languages, such as Czech, rely on overt inflectional and derivational morphology to convey many semantic distinctions. Languages with comparatively limited morphology, such as English, should be able to accomplish the same using a combination of syntactic and contextual cues. We capitalize on this idea by training a tagger for English that uses syntactic features obtained by automatic parsing to recover complex morphological tags projected from Czech. The high accuracy of the resulting model provides quantitative confirmation of the underlying linguistic hypothesis of equal expressivity, and bodes well for future improvements in downstream HLT tasks including machine translation.

1 Introduction

Different languages use different grammatical tools to convey the same meanings. For example, to indicate that a noun functions as a direct object, English—a morphologically poor language—places the noun after the verb, while Czech—a morphologically rich language—uses an accusative case suffix. Consider the following two glossed Czech sentences: *ryba jedla* ("the fish ate") and *oni jedli rybu* ("they ate the fish"). The key insight is that the morphology of Czech (i.e., the case ending *-u*), carries the same semantic content as the syntactic structure of English (i.e., the word order) (Harley, 2015). Theoretically, this common underlying semantics should allow syntactic structure to be transformed into morphological structure and vice versa. We explore the veracity of this claim computationally by asking the following: Can we develop a tagger for English that uses the signal available in English-only syntactic structure to recover the rich semantic distinctions conveyed by morphology in Czech? Can we, for example, accurately detect which English contexts would have a Czech translation that employs the accusative case marker?

Traditionally, morphological analysis and tagging is a task that has been limited to morphologically rich languages (MRLs) (Hajič, 2000; Drábek and Yarowsky, 2005; Müller et al., 2015; Buys and Botha, 2016). In order to build a rich morphological tagger for a morphologically poor language (MPL) like English, we need some way to build a gold standard set of richly tagged English data for training and testing. Our approach is to project the complex morphological tags of Czech words directly onto the English words they align to in a large parallel corpus. After evaluating the validity of these projections, we develop a neural network tagging architecture that takes as input a number of English features derived from off-the-shelf dependency parsing and attempts to recover the projected Czech tags.

A tagger of this sort is interesting in many ways. Whereas the best NLP tools are typically available for English, morphological tagging at this granularity has until now been applied almost exclusively to MRLs. The task is also scientifically interesting, in that it takes semantic properties that are latent in the syntactic structure of English and transforms them into explicit word-level annotations. Finally, such a tool has potential utility in a

Subtag	Values
GENDER	FEM, MASC, NEUT
NUMBER	SG, DU, PL
CASE	NOM, GEN, DAT, ACC, VOC, ESS, INS
PERSON	1, 2, 3
TENSE	FUT, PRS, PST
GRADE	CMPR, SPRL
NEGATION	POS, NEG
VOICE	ACT, PASS

Table 1: The subset of the UniMorph Schema used here.

range of downstream tasks, such as machine translation into MRLs (Sennrich and Haddow, 2016).

2 Projecting Morphological Tags

Training a system to tag English text with multi-dimensional morphological tags requires a corpus of English text annotated with those tags. Since no such corpora exist, we must construct one. Past work (focused on translating out of English into MRLs) assigned a handful of morphological annotations using manually-developed heuristics (Drábek and Yarowsky, 2005; Avramidis and Koehn, 2008), but this is hard to scale. We therefore instead look to obtain rich morphological tags by projecting them (Yarowsky et al., 2001) from a language (such as Czech) where such rich tags have already been annotated.

We use the Prague Czech–English Dependency Treebank (PCEDT) (Hajič et al., 2012), a complete translation of the Wall Street Journal portion of the Penn Treebank (PTB) (Marcus et al., 1993). Each word on the Czech side of the PCEDT was originally hand-annotated with complex 15-dimensional morphological tags containing positional subtag values for morphological categories specific to Czech.[1] We manually mapped these tags to the UniMorph Schema tagset (Sylak-Glassman et al., 2015), which provides a universal, typologically-informed annotation framework for representing morphological features of inflected words in the world's languages. UniMorph tags are in principle up to 23-dimensional, but tags are not positionally dependent, and not every dimension needs to be specified. Table 1 shows the subset of UniMorph subtags used here. PTB tags have no formal internal subtag structure.

PTB	Expected UM	Match %
NN	SG	87.8
NNP	SG	73.9
NNS	PL	83.3
NNPS	PL	65.1
JJR	CMPR	89.0
JJS	SPRL	79.3
RBR	CMPR	76.3
RBS	SPRL	68.7
VBZ	SG	91.3
VBZ	3	90.7
VBZ	PRS	89.4
VBG	PRS	55.9
VBP	PRS	87.2
VBD	PST	93.9
VBN	PST	78.7
Average Match %		80.7

Table 2: To evaluate the validity of projecting morphological tags from Czech onto English text, we compare these projected features to features obtained from the original PTB tags (listed on the left). The expected UniMorph (UM) subtag (center) is from a manual 'translation' of PTB tags into UniMorph tags. The match percentage indicates how often the feature projected from a UniMorph 'translation' of the original PCEDT annotation of Czech matches the feature that would be expected subtag. Note that the core part-of-speech must agree as a precondition for further evaluation.

See Figure 1 for a comparison of the PCEDT, UniMorph, and PTB tag systems for a Czech word and its aligned English translation.

The PCEDT also contains automatically generated word alignments produced by using GIZA++ (Och and Ney, 2003) to align the Czech and English sides of the treebank. We use these alignments to project morphological tags from the Czech words to their English counterparts through the following process. For every English word, if the word is aligned to a single Czech word, take its tag. If the word is mapped to multiple Czech words, take the annotation from the alignment point belonging to the intersection of the two underlying GIZA++ models used to produce the many-many alignment.[2] If no such alignment point is found, take the leftmost aligned word. Unaligned English words get no annotation.

3 Validating Projections

If we believe that we can project semantic distinctions over bitext, we must ensure that the elements linked by projection in both source and target languages carry roughly the same meaning. This is difficult to automate, and no gold-standard dataset or metric has been developed. Thus, we offer the following approximate evaluation.

[1] For our purposes, a morphological *tag* is a complex, multiclass entity comprising the morphological features that a word bears across many different inflectional categories (e.g., CASE, NUMBER, and so on). We call these features *subtags*, and each takes one of several values (e.g., PRS 'present' in the TENSE category of the UniMorph Schema).

[2] This intersection is marked as *int.gdfa* in the PCEDT.

Czech	PCEDT tag	UniMorph tag	=	English	PTB tag
je	VB-S---3P-AA---	V;ACT;POS;PRS;3;SG		is	VBZ

Figure 1: The PCEDT tag of the Czech word *je* was mapped to an equivalent UniMorph tag. The English translation of *je*, which is the copula *is*, has the PTB tag VBZ. While the PCEDT and UniMorph tags are composed of subtags, the PTB tag has no formal internal composition.

English is not bereft of morphological marking, and its use of it, though limited, does sometimes coincide with that of Czech. For example, both languages use overt morphology to mark nouns as *singular* or *plural*, adjectives and adverbs as *superlative* or *comparative*, and verbs as either *present* or *past*.[3] In these cases it is possible to directly map word-level PTB tags in English to word-level UniMorph tags in Czech, and to compare how often projected tags conform to this expected mapping. For example, the PTB tag VBZ is mapped to the UniMorph tag V;PRS;3;SG. Table 2 shows a set of expected projections along with how often the expectations are met across the PCEDT. In particular, we calculate the percentage of cases when an English word with a particular PTB tag has the expected Czech tag projected onto it. This calculation is only performed in those cases where where the aligned words agree in their core part of speech, since we would not expect, for example, verbs to have superlative/comparative morphology.

A qualitative examination of these results suggests that projections are usually valid in at least those cases where our limited linguistic intuitions predict they should be. For example, the dual number feature (DU) was projected in only 12 instances, but was almost always projected to the English words “two,” “eyes,” “feet,” and “hands.” These concepts naturally come in pairs, and this distinction is explicitly marked in Czech, but not English. We interpret this evaluation as suggesting that we can trust projection even in cases where we do not have pre-existing expectations of how English and Czech grammars should align.

4 Neural Morphological Tag Prediction

4.1 Features

With our projections validated, we turn to the prediction model itself. Based on the idea that languages with rich morphology use that morphology to convey similar distinctions in meaning to that conveyed by syntax in a morphologically poor language, we extract lexical and syntactic features from English text itself as well as both dependency and CFG parses. We use the following basic features derived directly from the text: the word itself, the single-word lexical context, and the word’s POS tag neighbors. We also use features derived from dependency trees.

- *Head features*. The word’s head word, and separately, the head word’s POS.
- *Head chain POS*. The chain of POS tags beginning with the word and moving upward along the dependency graph.
- *Head chain labels*. The chain of dependency labels moving upward.
- *Child words*. The identity of any child word having an arc label of *det* or *case*, under the Universal Dependency features.[4]

Finally, we use features from CFG parsing:

- *POS features*. A word’s part-of-speech (POS) tag, its parent’s, and its grandparent’s.
- *Chain features*. We compute chains of the tree nodes, starting with its POS tag and moving upward (*NN_NP_S*).
- The distance to the root.

Non-lexical features are treated as real-valued when appropriate (such as in the case of the distance to the root), while all others are treated as binary. For lexical features, we use pretrained GLoVe embeddings, specifically 200-dimensional 400K-vocab uncased embeddings from Pennington et al. (2014). This is an approach similar to Tran et al. (2015), but we additionally augment the pretrained embeddings with randomly initialized embeddings for vocabulary items outside of the 400K lexicon.

4.2 Neural Model

In order to take advantage of correlated information between subtags, we present a neural model

[3] English also uses morphology to mark the 3rd person singular verb form.

[4] `universaldependencies.org`

Other	*companies*	*are*	*introducing*	*related*	*products*
PL, NOM	PL, NOM	ACT, 3, PRS, PL	ACT, 3, PRS, PL	PL, ACC	PL, ACC

Table 3: An English sentence from the test set, WSJ §22, tagged with rich morphological tags by our neural tagger. Note, for example, that case is tagged correctly, with *Other companies* tagged as nominative and *related products* tagged as accusative. Legend here: CASE (NOM = nominative, ACC = accusative), TENSE (PRS = present), NUMBER (PL = plural), VOICE (ACT = active), and PERSON (3).

which learns a common representation of input tokens, and passes it on to a series of subtag classifiers that are trained jointly. Informally, this means that we learn a shared representation in the hidden layers and then use separate weight functions to predict each component of the morphological analysis from this shared representation of the input. We use a feed-forward neural net with two hidden layers and rectified linear unit (ReLU) activation functions (Glorot et al., 2011). A UniMorph tag m can be decomposed into its N subtags as $m = [m^{(1)}, m^{(2)}, \ldots, m^{(N)}]$, where each $m^{(i)}$ may be represented as a one-hot vector. The weight matrices ($W^{(1)}$, $W^{(2)}$) and bias vectors ($b^{(1)}$, $b^{(2)}$) connecting the hidden layers are parameters for the whole model, but each of the N subtag classes has its own weight matrix and bias vector $W_i^{(3)}, b_i^{(3)}$. All are randomly initialized from truncated normal distributions. Given an input vector x, we first compute a new input $x' = [x_{\text{non-lex}} : Ex_{\text{lex}_0} : Ex_{\text{lex}_1} : \ldots : Ex_{\text{lex}_n}]$, where $[a : b]$ represents vector concatenation. All lexical features x_{lex_i} are replaced by their embeddings from the embedding matrix E.

$$f(x')=\text{relu}\left(b^{(2)}+W^{(2)}\text{relu}\left(b^{(1)}+W^{(1)}x'\right)\right) \quad (1)$$

$$p(m^{(i)} \,|\, x, \theta) = \text{softmax}\left(b_i^{(3)} + W_i^{(3)} f(x')\right) \quad (2)$$

Then the definition of $p(m)$ follows:

$$p(m \mid x, \theta) = \prod_{i=1}^{N} p(m^{(i)} \mid x, \theta) \qquad (3)$$

The set of parameters is $\theta = \{E, W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}, W_1^{(3)}, b_1^{(3)}, \ldots, W_N^{(3)}, b_N^{(3)}\}$. The loss is defined as the cross-entropy, and the model is trained using gradient descent with minibatches. The models were trained using TensorFlow (Abadi et al., 2015). We complete a coarse-grained grid search over the learning rate, hidden layer size, and batch size. Based on performance on the development set, we choose a hidden layer size of 1000. We tune model parameters on whole-tag accuracy on WSJ §00. We find that a learning rate of 0.01 and batch size of 50 work best.

5 Experiment Setup

Our goal is to predict rich morphological tags for monolingual English text. The tagger was trained on §02–21 of the WSJ portion of the PTB. §00 was used for tuning. Training tags were projected from the equivalent Czech portion of the PCEDT, across the standard alignments provided by the PCEDT, as described in §2. Projected tags were treated as a gold standard to be recovered by the tagger. The full training set consisted of 39,832 sentences (726,262 words). Evaluation of the tagger was done on §22 of the WSJ portion of the PCEDT.

6 Results and Analysis

Table 4 shows the accuracy of the neural tagger for each subtag category from Table 1, indicating how often the tagger recovered the English projections of the Czech subtags. Baseline 1 is computed by selecting the most common Czech (sub)tag value in every case.

Baseline 2 is computed similarly to the evaluation of projection validity presented in §3. For each English word, the UniMorph subtag values which can be obtained by translating the PTB tag are compared to the projected subtag value in the same category (e.g. TENSE). This baseline penalizes cases in which a value for a category exists in the gold projection, but the value from the PTB tag translation either does not match or is not present at all. The poor performance of this baseline highlights how little information can be gleaned from traditional English PTB tags themselves, which is caused by the poverty of English inflectional morphology. In baselines 2 and 3, values for negation and voice were never present from the PTB tags since both negation and passive voice are indicated by separate words in English.[5]

[5] The tag VBN cannot be used in isolation to conclusively find use of the passive voice since it may occur in construc-

source	case	tense	per	num	neg	grade	voice	all
Baseline 1	35.0	86.7	94.2	45.6	68.8	99.0	86.7	14.1
Baseline 2	0.7	61.5	29.3	46.0	—	62.6	—	4.3
Baseline 3	46.4	89.1	99.8	86.3	—	99.5	—	8.6
PCEDT	**69.1**	**93.3**	*96.5*	*78.3*	*89.4*	**99.5**	*93.7*	**54.7**

Table 4: Performance of the neural tagger on §22 of the WSJ portion of the PTB. We report both subtag and whole tag accuracies. Baseline 1 simply outputs the most frequent subtag value. Baseline 2 outputs the subtag value that can be obtained from a human-annotated PTB tag with the gold subtag, and penalizes both values from the PTB tag that are either incorrect or missing. Baseline 3 does the same comparison, but penalizes only incorrect values, not those which are missing. Accuracy which exceeds or equals all baselines is bolded while that which exceeds only baselines 1 and 2 is italicized.

features	case	tense	person	num.	neg.	grade	voice
POS	46.4	91.2	95.3	68.7	84.2	99.3	91.8
Word	56.2	91.5	95.5	72.4	85.9	99.4	91.9
Word, POS	58.6	92.1	95.9	74.4	88.3	99.4	92.6
Word, POS, POS ctxt	63.8	92.7	96.1	77.5	89.1	**99.5**	93.2
CFG	65.0	92.7	96.2	77.5	88.8	99.4	93.1
dep	67.0	92.9	96.3	77.9	89.3	**99.5**	93.2
dep, CFG	**69.1**	92.9	96.4	78.0	89.2	**99.5**	93.2
dep, CFG, lex. ctxt	69.0	**93.2**	**96.6**	**79.1**	**89.8**	**99.5**	**93.7**

Table 5: Performance of the PCEDT-trained MaxEnt classifiers on §22 of the WSJ portion of the Penn Treebank. Bolding indicates the highest performance among the MaxEnt classifiers.

In baseline 3, we remove the effect of morphological poverty from consideration by comparing the values obtained from PTB tag translation to gold projected values only when both sources provide a value for a given category. The strong performance of this baseline, particularly in person and number, may be partly due to the fact that the tags are human-annotated as well as the fact that fewer comparisons are made in an attempt to isolate the effects of morphological poverty. In addition, baseline 3 need only predict instances of 3rd person, since person is only marked by PTB tags for one tag, VBZ. Similarly, PTB tags only explicitly mark number for the tags VBZ, NN, NNS, NNP, and NNPS.

The neural tagger outperforms baselines 1 and 2 everywhere, showing that the syntactic structure of English does contain enough signal to recover the complex semantic distinctions that are overt in Czech morphology. For case, especially, accuracy is nearly double that of baseline 1. Table 3 shows an example English sentence, where case and number have been tagged correctly. We examined the contribution of different grammatical aspects of English by training standard MaxEnt classifiers for each subtag using different subsets of features. The individual classifiers were trained with Liblinear's (Fan et al., 2008) MaxEnt model. We varied the regularization constant from 0.001 to 100 in multiples of 10, choosing the value in each situation that maximized performance on the dev set, PCEDT §00. Table 5 contains the results. First, word identity contributes more than POS on its own. This suggests that the distribution of morphological features is at least partially conditioned by lexical factors, in addition to grammatical properties such as POS. The addition of POS context, which includes the POS of the preceding and the following word, yields modest gains, except for case, in which it leads to a 5.2% increase in accuracy. POS context can be viewed as an approximation of true syntactic features, which yield greater improvements. Dependency parse features are particularly effective in helping to predict case since case is typically assigned by a verb governing a noun in a head-dependency relationship. The direct encoding of this relationship yields an especially salient feature for the case classifier. Even with these improvements, the case feature remains the most difficult to predict, suggesting that even more salient features have yet to be discovered.

7 Conclusion

To our knowledge, this is the first work to construct a rich morphological tagger for English that does not rely on manually-developed syntactic heuristics. This significantly extends the applicability and usability of the proposed general tagging framework, which offers the ability to use automatic parsing features in one language and (potentially automatically generated) morphological feature annotation in the other. Validating the claim that languages apply different aspects of grammar to represent equivalent meanings, we find that English-only lexical, contextual, and syntactic features derived from off-the-shelf parsing tools encode the complex semantic distinctions present in Czech morphology. In addition to allowing this scientific claim to be computationally validated, we expect this approach to generalize to tagging any morphologically poor language with the morphological distinctions made in another morphologically rich language.

tions such as 'have given' in which the VP as a whole is not passive.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 763–770, Columbus, Ohio, June. Association for Computational Linguistics.

Jan Buys and Jan A. Botha. 2016. Cross-lingual morphological tagging for low-resource languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1954–1964, Berlin, August. Association for Computational Linguistics.

Elliott Franco Drábek and David Yarowsky. 2005. Induction of fine-grained part-of-speech taggers via classifier combination and crosslingual projection. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 49–56, Ann Arbor, June. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 315–323.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing prague czech-english dependency treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey. ELRA, European Language Resources Association.

Jan Hajič. 2000. Morphological tagging: Data vs. dictionaries. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics Conference (NAACL 2000)*, pages 94–101, Seattle, May. Association for Computational Linguistics.

Heidi Harley. 2015. The syntax-morphology interface. In Tibor Kiss and Artemis Alexiadou, editors, *Syntax - Theory and Analysis: An International Handbook*, volume II, pages 1128–1153. Mouton de Gruyter, Berlin.

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

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal, September. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the 1st Conference on Machine Translation*, volume 1, pages 83–91, Berlin, August. Association for Computational Linguistics.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. 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 674–680, Beijing, China, July. Association for Computational Linguistics.

Ke Tran, Arianna Bisazza, and Christof Monz. 2015. A distributed inflection model for translating into morphologically rich languages. In *Proceedings of MT-Summit 2015*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT '01 Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8, Stroudsburg, PA. Association for Computational Linguistics.