A Universal Part-of-Speech Tagset

Slav Petrov
Google Research
New York, NY, USA
slav@google.com

Dipanjan Das
Carnegie Mellon University
Pittsburgh, PA, USA
dipanjan@cs.cmu.edu

Ryan McDonald
Google Research
New York, NY, USA
ryanmcd@google.com

Abstract

To facilitate future research in unsupervised induction of syntactic structure and to standardize best-practices, we propose a tagset that consists of twelve universal part-of-speech categories. In addition to the tagset, we develop a mapping from 25 different treebank tagsets to this universal set. As a result, when combined with the original treebank data, this universal tagset and mapping produce a dataset consisting of common parts-of-speech for 22 different languages. We highlight the use of this resource via two experiments, including one that reports competitive accuracies for unsupervised grammar induction without gold standard part-of-speech tags.

1 Introduction

Part-of-speech (POS) tagging has received a great deal of attention as it is a critical component of most natural language processing systems. As supervised POS tagging accuracies for English (measured on the Wall Street Journal portion of the PennTreebank (Marcus et al., 1993)) have converged to around 97.3% (Toutanova et al., 2003; Shen et al., 2007), the attention has shifted to unsupervised approaches (Christodoulopoulos et al., 2010). In particular, there has been growing interest in both multilingual POS induction (Snyder et al., 2009; Naseem et al., 2009) and cross-lingual POS induction via treebank projection (Yarowsky and Ngai, 2001; Xi and Hwa, 2005; Das and Petrov, 2011).

Underlying these studies is the idea that a set of (coarse) syntactic POS categories exist in similar forms across languages. These categories are often called *universals* to represent their cross-lingual nature (Carnie, 2002; Newmeyer, 2005). For example, Naseem et al. (2009) used the Multext-East (Erjavec, 2004) corpus to evaluate their multi-lingual POS induction system, because it uses the same tagset for multiple languages. When corpora with common tagsets are unavailable, a standard approach is to manually define a mapping from language and treebank specific fine-grained tagsets to a predefined universal set. This was the approach taken by Das and Petrov (2011) to evaluate their cross-lingual POS projection system for six different languages.

To facilitate future research and to standardize best-practices, we propose a tagset that consists of twelve universal POS categories. While there might be some controversy about what the exact tagset should be, we feel that these twelve categories cover the most frequent part-of-speech that exist in most languages. In addition to the tagset, we also develop a mapping from fine-grained POS tags for 25 different treebanks to this universal set. As a result, when combined with the original treebank data, this universal tagset and mapping produce a dataset consisting of common parts-of-speech for 22 different languages.[1] Both the tagset and mappings are made available for download at **http://code.google.com/p/universal-pos-tags/**.

This resource serves multiple purposes. First, as mentioned previously, it is useful for building and evaluating unsupervised and cross-lingual taggers. Second, it also permits for a more reasonable com-

[1] We include mappings for two different Chinese, German and Japanese treebanks.

sentence:	The	oboist	Heinz	Holliger	has	taken	a	hard	line	about	the	problems	.
original:	DT	NN	NNP	NNP	VBZ	VBN	DT	JJ	NN	IN	DT	NNS	.
universal:	DET	NOUN	NOUN	NOUN	VERB	VERB	DET	ADJ	NOUN	ADP	DET	NOUN	.

Figure 1: Example English sentence with its language specific and corresponding universal POS tags.

parison of accuracy across languages for supervised taggers. Statements of the form "POS tagging for language X is harder than for language Y" are vacuous when the tagsets used for the two languages are incomparable (not to mention of different cardinality). Finally, it also permits language technology practitioners to train POS taggers with common tagsets across multiple languages. This in turn facilitates downstream application development as there is no need to maintain language specific rules due to differences in treebank annotation guidelines.

In this paper, we specifically highlight two use cases of this resource. First, using our universal tagset and mapping, we run an experiment comparing POS tag accuracies for 25 different treebanks to evaluate POS tagging accuracy on a single tagset. Second, we combine the cross-lingual projection part-of-speech taggers of Das and Petrov (2011) with the grammar induction system of Naseem et al. (2010) – which requires a universal tagset – to produce a completely unsupervised grammar induction system for multiple languages, that does not require gold POS tags in the target language.

2 Tagset

While there might be some disagreement about the exact definition of an universal POS tagset (Evans and Levinson, 2009), it seems fairly indisputable that a set of coarse POS categories (or syntactic universals) exists across all languages in one form or another (Carnie, 2002; Newmeyer, 2005). Rather than arguing over definitions, we took a pragmatic standpoint during the design of the universal POS tagset and focused our attention on the POS categories that we expect to be most useful (and necessary) for users of POS taggers. In our opinion, these are NLP practitioners using POS taggers in downstream applications, and NLP researchers using POS taggers in grammar induction and other experiments.

A high-level analysis of the tagsets underlying various treebanks shows that the majority of tagsets are very fine-grained and language specific. In fact, Smith and Eisner (2005) made a similar observation and defined a collapsed set of 17 English POS tags (instead of the original 45) that has subsequently been adopted by most unsupervised POS induction work. Similarly, the organizers of the CoNLL shared tasks on dependency parsing provide coarse (but still language specific) tags in addition to the fine-grained tags used in the original treebanks (Buchholz and Marsi, 2006; Nivre et al., 2007). McDonald and Nivre (2007) identified eight different coarse POS tags when analyzing the errors of two dependency parsers on the 13 different languages from the CoNLL shared tasks.

Our universal POS tagset unifies this previous work and defines the following twelve POS tags: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (a catch-all for other categories such as abbreviations or foreign words).

We did not rely on intrinsic definitions of the above categories. Instead, each category is defined operationally. For each treebank under consideration, we studied the exact POS tag definitions and annotation guidelines and created a mapping from the original treebank tagset to these universal POS tags. Most of the decisions were fairly clear. For example, from the PennTreebank, VB, VBD, VBG, VBN, VBP, VBZ and MD (modal) were all mapped to VERB. A less clear case was the universal tag for particles, PRT, which was mapped from POS (possessive), RP (particle) and TO (the word 'to'). In particular, the TO tag is ambiguous in the PennTreebank between infinitival markers and the preposition 'to'. Thus, under this mapping, some prepositions will be marked as particles in the universal tagset. Figure 1 gives an example mapping for a sentence from the PennTreebank.

Another case we had to consider is that some tag

Language	Source	# Tags	O/O	U/U	O/U
Arabic	PADT/CoNLL07 (Hajič et al., 2004)	21	96.1	96.9	97.0
Basque	Basque3LB/CoNLL07 (Aduriz et al., 2003)	64	89.3	93.7	93.7
Bulgarian	BTB/CoNLL06 (Simov et al., 2002)	54	95.7	97.5	97.8
Catalan	CESS-ECE/CoNLL07 (Martí et al., 2007)	54	98.5	98.2	98.8
Chinese	Penn ChineseTreebank 6.0 (Palmer et al., 2007)	34	91.7	93.4	94.1
Chinese	Sinica/CoNLL07 (Chen et al., 2003)	294	87.5	91.8	92.6
Czech	PDT/CoNLL07 (Böhmová et al., 2003)	63	99.1	99.1	99.1
Danish	DDT/CoNLL06 (Kromann et al., 2003)	25	96.2	96.4	96.9
Dutch	Alpino/CoNLL06 (Van der Beek et al., 2002)	12	93.0	95.0	95.0
English	PennTreebank (Marcus et al., 1993)	45	96.7	96.8	97.7
French	FrenchTreebank (Abeillé et al., 2003)	30	96.6	96.7	97.3
German	Tiger/CoNLL06 (Brants et al., 2002)	54	97.9	98.1	98.8
German	Negra (Skut et al., 1997)	54	96.9	97.9	98.6
Greek	GDT/CoNLL07 (Prokopidis et al., 2005)	38	97.2	97.5	97.8
Hungarian	Szeged/CoNLL07 (Csendes et al., 2005)	43	94.5	95.6	95.8
Italian	ISST/CoNLL07 (Montemagni et al., 2003)	28	94.9	95.8	95.8
Japanese	Verbmobil/CoNLL06 (Kawata and Bartels, 2000)	80	98.3	98.0	99.1
Japanese	Kyoto4.0 (Kurohashi and Nagao, 1997)	42	97.4	98.7	99.3
Korean	Sejong (http://www.sejong.or.kr)	187	96.5	97.5	98.4
Portuguese	Floresta Sintá(c)tica/CoNLL06 (Afonso et al., 2002)	22	96.9	96.8	97.4
Russian	SynTagRus-RNC (Boguslavsky et al., 2002)	11	96.8	96.8	96.8
Slovene	SDT/CoNLL06 (Džeroski et al., 2006)	29	94.7	94.6	95.3
Spanish	Ancora-Cast3LB/CoNLL06 (Civit and Martí, 2004)	47	96.3	96.3	96.9
Swedish	Talbanken05/CoNLL06 (Nivre et al., 2006)	41	93.6	94.7	95.1
Turkish	METU-Sabanci/CoNLL07 (Oflazer et al., 2003)	31	87.5	89.1	90.2

Table 1: Data sets, number of language specific tags in the original treebank, and tagging accuracies for training/testing on the original (O) and the universal (U) tagset. Where applicable, we indicate whether the data set was extracted from the CoNLL 2006 (Buchholz and Marsi, 2006) or CoNLL 2007 (Nivre et al., 2007) versions of the corpora.

categories do not occur in all languages. Consider for example the case of adjectives. While all languages have a way of describing the properties of objects (which themselves are typically referred to with nouns), many have argued that Korean does not technically have adjectives, but instead expresses properties of nouns via stative verbs (Kim, 2002). As a result, in our mapping for Korean, we mapped stative verbs to the universal ADJ tag. In other cases this was clearer, e.g. the Bulgarian treebank has no category for determiners or articles. This is not to say that there are no determiners in the Bulgarian language. However, since they are not annotated as such in the treebank, we are not able to include them in our mapping.

The list of treebanks for which we have constructed mappings can be seen in Table 1. One main objective in publicly releasing this resource is to provide treebank and language specific experts a mechanism for refining these categories and the decisions we have made, as well as adding new treebanks and languages. This resource is therefore hosted as an open source project with version control.

3 Experiments

To demonstrate the efficacy of the proposed universal POS tagset, we performed two sets of experiments. First, to provide a language comparison, we trained the same supervised POS tagging model on all of the above treebanks and evaluated the tagging accuracy on the universal POS tagset. Second, we used universal POS tags (automatically projected from English) as the starting point for unsupervised grammar induction, producing completely unsupervised parsers for several languages.

3.1 Language Comparisons

To compare POS tagging accuracies across different languages we trained a supervised tagger based on a trigram Markov model (Brants, 2000) on all tree-

banks. We chose this model for its fast speed and (close to) state-of-the-art accuracy without language specific tuning.[2]

Table 1 shows the results for all 25 treebanks when training/testing on the original (O) and universal (U) tagsets. Overall, the variance on the universal tagset has been reduced by half (5.1 instead of 10.4). But of course there are still accuracy differences across the different languages. On the one hand, given a golden segmentation, tagging Japanese is almost deterministic, resulting in a final accuracy of above 99%.[3] On the other hand, tagging Turkish, an agglutinative language with and average sentence length of 11.6 tokens, remains very challenging, resulting in an accuracy of only 90.2%.

It should be noted that the best results are obtained by training on the original treebank categories and mapping the predictions to the universal POS tags at the end (O/U column). This is because the transition model based on the universal POS tagset is less informative. An interesting experiment would be to train the latent variable tagger of Huang et al. (2009) on this tagset. Their model automatically discovers refinements of the observed categories and could potentially find a tighter fit to the data, than the one provided by the original, linguistically motivated treebank tags.

3.2 Grammar Induction

We further demonstrate the utility of the universal POS tags in a grammar induction experiment. To decouple the challenges of POS tagging and parsing, golden POS tags are typically assumed in unsupervised grammar induction experiments (Carroll and Charniak, 1992; Klein and Manning, 2004).[4] We propose to remove this unrealistic simplification by using POS tags automatically projected from English as the basis of a grammar induction model.

Das and Petrov (2011) describe a cross-lingual projection framework to learn POS taggers without labeled data for the language of interest. We use their automatically induced POS tags to induce syntactic dependencies. To this end, we chose the framework of Naseem et al. (2010), in which a few universal syntactic rules (USR) are used to constrain a probabilistic Bayesian model. These rules are specified using a set of universal syntactic categories, and lead to state-of-the-art grammar induction performance superior to previous methods, such as the dependency model with valence (DMV) (Klein and Manning, 2004) and the phylogenetic grammar induction model (PGI) (Berg-Kirkpatrick and Klein, 2010).

Language	DMV	PGI	USR-G	USR-I
Danish	33.5	41.6	55.1	41.7
Dutch	37.1	45.1	44.0	38.8
German	35.7	-[5]	60.0	55.1
Greek	39.9	-[5]	60.3	53.4
Italian	41.1	-[5]	47.9	41.4
Portuguese	38.5	63.0	70.9	66.4
Spanish	28.0	58.4	68.3	43.3
Swedish	45.3	58.3	52.6	59.4

Table 2: Grammar induction results in terms of directed dependency accuracy. DMV, PGI and use fine-grained gold POS tags, while USR-G and USR-I uses gold and automatically projected universal POS tags respectively.

In their experiments, Naseem et al. also used a set of universal categories, however, with some differences to the tagset presented here. Their tagset does not have punctuation and catch-all categories, but includes a category for auxiliaries. The auxiliary category helps define a syntactic rule that attaches verbs to an auxiliary head, which is beneficial for certain languages. However, since this rule is reversed for other languages, we omit it in our tagset. Additionally, they also used refined categories in the form of CoNLL treebank tags. In our experiments, we did not make use of refined categories, as the POS tags induced by Das and Petrov (2011) were all coarse.

We present results on the same eight Indo-European languages as Das and Petrov (2011), so that we can make use of their automatically projected POS tags. For all languages, we used the treebanks released as a part of the CoNLL-X (Buchholz and Marsi, 2006) shared task. We only considered sentences of length 10 or less, after the removal of punctuations. We performed Bayesian inference on

[2] Trained on the English PennTreebank this model achieves 96.7% accuracy when evaluated on the original 45 POS tags.

[3] Note that the accuracy on the universal POS tags for the two Japanese treebanks is almost the same.

[4] A less benevolent explanation for this practice is that grammar induction from plain text is simply still too difficult.

[5] Not reported by Berg-Kirkpatrick and Klein (2010).

the whole treebank and report dependency attachment accuracy.

Table 2 shows directed dependency accuracies for the DMV and PGI models using fine-grained gold POS tags. For the USR model, it reports results on gold universal POS tags (USR-G) and automatically induced universal POS tags (USR-I). The USR-I model falls short of the USR-G model, but has the advantage that it does not require any labeled data from the target language. Quite impressively, it does better than DMV for all languages, and is competitive with PGI, even though those models have access to fine-grained gold POS tags.

4 Conclusions

We proposed a POS tagset consisting of twelve categories that exists across languages and developed a mapping from 25 language specific tagsets to this universal set. We demonstrated experimentally that the universal POS categories generalize well across language boundaries on an unsupervised grammar induction task, giving competitive parsing accuracies without relying on gold POS tags. The tagset and mappings are available for download at **http://code.google.com/p/universal-pos-tags/**

Acknowledgements

We would like to thank Joakim Nivre for allowing us to use a preliminary tagset mapping used in the work of McDonald and Nivre (2007). The second author was supported in part by NSF grant IIS-0844507.

References

A. Abeillé, L. Clément, and F. Toussenel. 2003. Building a Treebank for French. In Abeillé (Abeillé, 2003), chapter 10.

A. Abeillé, editor. 2003. *Treebanks: Building and Using Parsed Corpora*. Kluwer.

I. Aduriz, M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. Diaz de Ilarraza, A. Garmendia, and M. Oronoz. 2003. Construction of a Basque dependency treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*.

S. Afonso, E. Bick, R. Haber, and D. Santos. 2002. Floresta sintá(c)tica: a treebank for Portuguese. In *Proc. of LREC*.

T. Berg-Kirkpatrick and D. Klein. 2010. Phylogenetic grammar induction. In *Proc. of ACL*.

I.M. Boguslavsky, L.L. Iomdin, I.S. Chardin, and L.G. Kreidlin. 2002. Development of a dependency treebank for russian and its possible applications in nlp. In *Proc. of LREC*.

A. Böhmová, J. Hajič, E. Hajičová, and B. Hladká. 2003. The PDT: a 3-level annotation scenario. In Abeillé (Abeillé, 2003), chapter 7, pages 103–127.

S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The TIGER Treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*.

T. Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proc. of ANLP*.

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proc. of CoNLL*.

A. Carnie. 2002. *Syntax: A Generative Introduction (Introducing Linguistics)*. Blackwell Publishing.

G. Carroll and E. Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. In *Working Notes of the Workshop Statistically-Based NLP Techniques*.

K. Chen, C. Luo, M. Chang, F. Chen, C. Chen, C. Huang, and Z. Gao. 2003. Sinica treebank: Design criteria, representational issues and implementation. In Abeillé (Abeillé, 2003), chapter 13, pages 231–248.

C. Christodoulopoulos, S. Goldwater, and M. Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proc. of EMNLP*.

M. Civit and M.A. Martí. 2004. Building cast3lb: A spanish treebank. *Research on Language & Computation*, 2(4):549–574.

D. Csendes, J. Csirik, T. Gyimóthy, and A. Kocsor. 2005. *The Szeged Treebank*. Springer.

D. Das and S. Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. of ACL-HLT*.

S. Džeroski, T. Erjavec, N. Ledinek, P. Pajas, Z. Žabokrtsky, and A. Žele. 2006. Towards a Slovene dependency treebank. In *Proc. of LREC*.

T. Erjavec. 2004. MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc. of LREC*.

N. Evans and S. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(05).

J. Hajič, O. Smrž, P. Zemánek, J. Šnaidauf, and E. Beška. 2004. Prague Arabic dependency treebank: Development in data and tools. In *Proc. of NEMLAR*.

Z. Huang, V. Eidelman, and M. Harper. 2009. Improving simple bigram HMM part-of-speech tagger by latent annotation. In *Proc. of NAACL-HLT*.

Y. Kawata and J. Bartels. 2000. Stylebook for the Japanese treebank in VERBMOBIL.

M.J. Kim. 2002. Does Korean have adjectives? *MIT Working Papers in Linguistics*, 43:71–89.

D. Klein and C. D. Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *Proc. of ACL*.

M.T. Kromann, L. Mikkelsen, and S.K. Lynge. 2003. Danish Dependency Treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*.

S. Kurohashi and M. Nagao. 1997. Kyoto University text corpus project. In *Proc. of ANLP*.

M. P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19.

M. A. Martí, M. Taulé, L. Màrquez, and M. Bertran. 2007. CESS-ECE: A multilingual and multilevel annotated corpus. Available for download from: http://www.lsi.upc.edu/∼mbertran/cess-ece/.

R. McDonald and J. Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proc. of EMNLP-CoNLL*.

S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Nana, F. Pianesi, and R. Delmonte. 2003. Building the Italian Syntactic-Semantic Treebank. In Abeillé (Abeillé, 2003), chapter 11, pages 189–210.

T. Naseem, B. Snyder, J. Eisenstein, and R. Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *JAIR*, 36.

T. Naseem, H. Chen, R. Barzilay, and M. Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proc. of EMNLP*.

F. J. Newmeyer. 2005. *Possible and Probable Languages: A Generative Perspective on Linguistic Typology.* Oxford University Press.

J. Nivre, J. Nilsson, and J. Hall. 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proc. of LREC*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proc. EMNLP-CoNLL*.

K. Oflazer, B. Say, D. Zeynep Hakkani-Tür, and G. Tür. 2003. Building a Turkish treebank. In Abeillé (Abeillé, 2003), chapter 15, pages 261–277.

M. Palmer, N. Xue, F. Xia, F. Chiou, Z. Jiang, and M. Chang. 2007. Chinese Treebank 6.0. Technical report, Linguistic Data Consortium, Philadelphia.

P. Prokopidis, E. Desypri, M. Koutsombogera, H. Papageorgiou, and S. Piperidis. 2005. Theoretical and practical issues in the construction of a Greek dependency treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*.

L. Shen, G. Satta, and A. Joshi. 2007. Guided learning for bidirectional sequence classification. In *Proc. of ACL*.

K. Simov, P. Osenova, S. Kolkovska, E. Balabanova, D. Doikoff, K. Ivanova, A. Simov, and M Kouylekov. 2002. Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. In *Proc. of LREC*.

W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proc. of ANLP*.

N. A. Smith and J. Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proc. of ACL*.

B. Snyder, T. Naseem, J. Eisenstein, and R. Barzilay. 2009. Adding more languages improves unsupervised multilingual part-of-speech tagging: A Bayesian nonparametric approach. In *Proc. of NAACL*.

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of HLT-NAACL*.

L. Van der Beek, G. Bouma, R. Malouf, and G. Van Noord. 2002. The Alpino dependency treebank. *Language and Computers*, 45(1):8–22.

C. Xi and R. Hwa. 2005. A backoff model for bootstrapping resources for non-English languages. In *Proc. of HLT-EMNLP*.

D. Yarowsky and G. Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proc. of NAACL*.