

# NLP Takeaways

---

DHEERAJ JOSHI



# The Motto “Getting Busy Living”


---

- ✓ To understand the complexity of tagging speeches in any language
- ✓ Explore common linguistic phenomenon across several languages to derive a universal POS tagging set.

# The “Work on the Field”

---

Observations upon studying several approaches of tagging the Brown Corpus were:

- ✓ Frequency of title case words is a better feature than case-sensitivity.
  - ✓ Filtering out frequent words with unambiguous tags reduces computational overhead.
  - ✓ Greedy Taggers eliminate a great amount of linguistic context.
- 

# First Brainstorm “102 Bookmarks”

---

Averaged Perceptron Approach:

- A set of features along with an associated weight for each feature.
- Receive (A set of features and a POS tag) pair.
- Using the current weights of these features, we predict the appropriate POS tag.
- To learn from the mistakes we make,
  - Increment the weights of the predicted class. (Increase Confidence)
  - Penalize weights that lead to false prediction. (Reduce Error)

Example “from I am a Tea-pot, fat and stout”

---

I	am	a	Keyboard	Font	Installer.
(PRP)	(VBP)	(DT)	(NN)	(NNP)	?

Pairs: {(Suffix[i-1],0.03), (Context[i-1],0.04).....

Result: Prediction 1: NNP (Weight + 1) 👍

Prediction 2: NN (Weight - 1) 👎

# Weight Updation “Just for Showing Off”

---

```
def update(self, truth, guess, features):  
    def upd_feat(c, f, v):  
        nr_iters_at_this_weight = self.i - self._timestamps[f][c]  
        self._totals[f][c] += nr_iters_at_this_weight * self.weights[f][c]  
        self.weights[f][c] += v  
        self._timestamps[f][c] = self.i  
  
    self.i += 1  
    for f in features:  
        upd_feat(truth, f, 1.0)  
        upd_feat(guess, f, -1.0)
```

## Second Thought “Wait a minute”

---

- ❑ Read about multi-lingual POS induction and Cross lingual induction using tree bank tagsets.
- ❑ Mapping Tree bank tagsets into a common universal set of POS tags standardizes the best practices in NLP and helps to induct the syntactic structure of several languages.

# Universal POS Tagging “Thanks Prof.”

---

*Paper by Das, Petrov & McDonald*

Repo: <https://github.com/slavpetrov/universal-pos-tags>

## ***Highlights:***

- 25 Languages Mapped
- Very good accuracies of Training vs. Testing datasets
- Unsupervised Grammar Induction produces efficient parsers.



# Results “Now I do believe”

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 Chinese Treebank 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	Penn Treebank (Marcus et al., 1993)	45	96.7	96.8	97.7
French	French Treebank (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 ( <a href="http://www.sejong.or.kr">http://www.sejong.or.kr</a> )	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-Sabancı/CoNLL07 (Ofłazer et al., 2003)	31	87.5	89.1	90.2

# The IDEA “maybe too far fetched”

---

- ✓ To recognize speech using a microphone & PySpeech API
- ✓ Universally POS tag the sentences in any of the 25 languages using the mappings available.

In your face, Class Next Door!

# References “You’re the Real MVP”

---

- <https://spacy.io/blog/part-of-speech-pos-tagger-in-python>
- <https://pythonspot.com/en/speech-recognition-using-google-speech-api/>
- <http://petrovi.de/data/universal.pdf>
- <http://www.nltk.org/book/ch03.html>