Role of surface morphological features in the induction of part-of-speech categories

A comparative study

Daniel Devatman Hromada[12]

[1]Slovak University of Technology
Faculty of Electronic Engeneering and Informatics
Department of Robotics and Cybernetics

[2]Université Paris 8
École Doctoralle Cognition, Langage, Interaction
Laboratoire Cognition Humaine et Artificielle

17th International Conference on Text, Speech and Dialogue,
Brno, Czech Republic, EU, 10.9.2014

Table of Contents

What SHALL & SHAN'T be presented

What shall NOT be presented

1. a highly performant English-focused POS-i system
2. dozens incomprehensible formulas and analytical proofs
3. a model which is useless beyond the domain for which it was created

What SHALL be presented

1. simple and swift system of part-of-speech induction
2. comparativist multi-lingual perspective
3. one interesting correlation
4. and one potentially universal principle of human linguistic development

Table of Contents

Part-of-speech induction

Definition

Unsupervised learning of grammatical (part-of-speech) categories from unlabeled text.

INPUT: text composed of sequences of tokens (i.e. words)
OUTPUT: part-of-speech categories (i.e. verbs, nouns, prepositions) attributed to all tokens of the text

Why POS-i matters

1. key to grammar induction and language acquisition
2. form of category construction problem
3. less ambiguous and closer to surface than semantic category induction

Existing models of Part-of-speech induction

- Markov chains, mutual information, expectation-maximization, variational Bayes, Gibbs sampling etc...
- Elman, 1989; Brown, 1992; Clark, 2003; Berg-Kirkpatrick et al., 2010
- and dozens of others...
- practically all based upon word n-gram conditional probability paradigm
- (nice comparative study in Christodoulopoulos et al., 2010)

Evaluation metrics

- external (i.e. with K golden standard classes) clustering evaluation techniques
- purity, Rand index, Jaccard index, Mutual information, Variation of information, F-measure, etc.
- V-measure (harmonic mean of homogenity and completeness)
  - homogenity: 1 when all clusters contain datapoints which are members of single class
  - completeness=1 all data points of a given class are elements of the same cluster
  - based on entropies of classes (H(C)), clusters (H(K)) and conditional entropies H(C | K) and H(K | C)

Table of Contents

Word-order

POS-i of target token is determined by its history (i.e. sequence of preceding tokens) or its neighbors within the sliding window of size S.

- based on co-occurrence, juxtaposition and conditional probabilities
- circular (in sequence ABCD, $POS_B \sim POS_A + POS_C$ and $POS_C \sim POS_B + POS_D$)
- particularly important in languages with poor morphology (like English)

Chicken and egg problem of human language acquisition

Does POS-knowledge preceeds or follows knowledge about syntactic rules and dependencies?

Surface features 1

Length [L] – yields only one feature whose value equals the character length of the token, i.e. 6 for word "^good$". Baseline.

Character n-grams of length X [N_x] – every feature encodes the number of occurrences of the character n-gram of length L within the token. Thus, if X=1, the word "^ good$" can be encoded by vector of features [1, 1, 2, 1, 1] whose second element denotes the number of "g" present in the word, third element the number of "o" etc.

Character fragments whose length <X [F_X] – this approach takes into account all n-gram fragments BELOW the specified length X. Thus if X=3, the word "^good$" could be represented by the vector [1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

All fragments [A] – same as above but X is equal to word's length.

Surface features 2

Prefixes of length X [P_X]– same as N_X but fragments of length X were extracted only from word's beginning.

Suffixes of length L [S_X]– same as P_X but fragments of length L were extracted only from word's end.

Word's circumference [C_X] – Union of P_X and S_X,i.e. token is characterized by both its prefixes and suffixes of length X.

Word's root [R_X]– all that rests in the token when its circumference n-grams of length X are removed.

Morphological features

- Syllables (accentuated, non-accentuated)
- True meaning-carrying morphemes (roots, prefixes, suffixes)
- Repetition-based (anaphoras, reduplication, retriplication)
- Alternations (apophony, ablaut, umlaut, sandhi...)
- etc.

- not addressed in this study

Table of Contents

Multext-East

All analyses were effectuated with texts contained in the 4th version of Multext-East corpus (Erjavec, 2012). 11 diverse translations of Orwell's 1984 were thus analysed.

Corpus	Types	Tokens	Tags_{POS}
bg	17305	117238	13
cs	22341	100368	13
en	11160	134832	12
et	18911	111305	12
fa	13009	124823	12
hu	20642	132196	13
pl	24019	115185	14
ro	16220	135055	15
sr	21540	126611	13
sk	23015	103452	13
sl	20597	112278	13

Processing

Pre-processing

All tokens, punctuation marks included, were extracted as such from the corpus. Word characters were transcribed into lower case. In order to mark the word boundaries, ^ and $ characters were prefixed, respectively suffixed, to extracted tokens.

Clustering

We used Simple (& fast) repeated bisection k-way clustering as is implemented in the clustered tool CLUTO (Karypis, 2002). Columns of the word x feature matrix were scaled according to inverse-document frequency paradigm, cosine function was used for the calculation of the similarity metrics.

V-measure Evaluation (Rosenberg et al., 2007)

$$h = 1 - \frac{H(T|C)}{H(T)}$$

$$c = 1 - \frac{H(C|T)}{H(C)}$$

$$V = \frac{(1+\beta)hc}{(\beta h) + c}$$

Advantages of VM

- multi-class variant of F-score
- stable in regards to number of clusters
- quite strict (state-of-the-art algos <0.6)
- values in <0, 1>interval and other nice properties

Results 1

	L	N_1	N_2	N_3	N_4	F_2	F_3	F_4	A
bg	4.3	5.6	13.1	17.0	11.9	8.5	14.4	14.7	14.6
cs	5.4	9.2	**25.2**	20.7	11.6	23.1	24.8	23.9	24.3
en	3.8	6.5	14.1	15.3	9.4	10.4	14.9	16.1	14.7
et	4.2	4.0	12.2	14.2	11.9	5.8	6.92	9.38	7.24
fa	2.6	6.8	15.4	15.5	12.2	12.0	15.51	15.3	15.55
hu	2.3	4.3	6.1	10.7	9.4	5.2	6.26	6.58	5.65
pl	4.7	8.0	21.1	20.1	13.7	18.5	20.3	19.7	15.6
ro	4.6	7.1	11.1	13.6	9.5	8.23	11.3	11.8	10.9
sr	5.2	5.5	13.3	14.8	10.5	5.67	8.06	8.82	5.95
sk	5.9	11.2	26.9	21.0	14.0	23.8	24.9	24.2	22.5
sl	4.5	4.8	12.2	17.1	12.8	7.39	8.42	14.3	7.5

Length [L]; Character n-grams of length X [N_x]; Character fragments whose length $<$X [F_X]; All fragments [A];

Results 2

	P_2	P_3	S_2	S_3	C_2	C_3	R_2	R_3	O_1
bg	6.7	5.0	**18.9**	16.5	3.8	2.3	3.4	3.0	12.5
cs	7.4	7.1	**25.2**	18.7	4.7	3.1	3.7	3.4	7.9
en	3.9	3.6	**20.5**	19.7	2.4	1.7	2.9	2.2	14.4
et	4.2	6.0	14.2	**16.1**	3.6	2.8	3.4	3.3	6.77
fa	11.7	14.5	14.4	12.0	6.4	4.6	2.8	3.2	14.3
hu	5.4	5.7	**17.1**	14.2	3.0	1.8	2.4	2.0	7.1
pl	5.3	6.5	**25.1**	22.7	4.0	3.0	3.3	2.9	7.9
ro	6.5	5.9	**15.8**	14.8	3.1	1.9	2.5	2.4	15.6
sr	6.1	6.4	**19.1**	16.5	4.6	3.0	4.7	3.5	9.4
sk	8.2	5.8	**27.5**	21.3	4.8	3.5	3.6	3.5	8.7
sl	6.8	6.0	**21.6**	19.3	5.2	2.4	3.3	3.4	9.1

Prefixes of length X [P_X]; Suffixes of length X [S_X]; Circumference [C_X]; "Root" [R_X]; Neighborhing tokens [O_1]

Comparative results

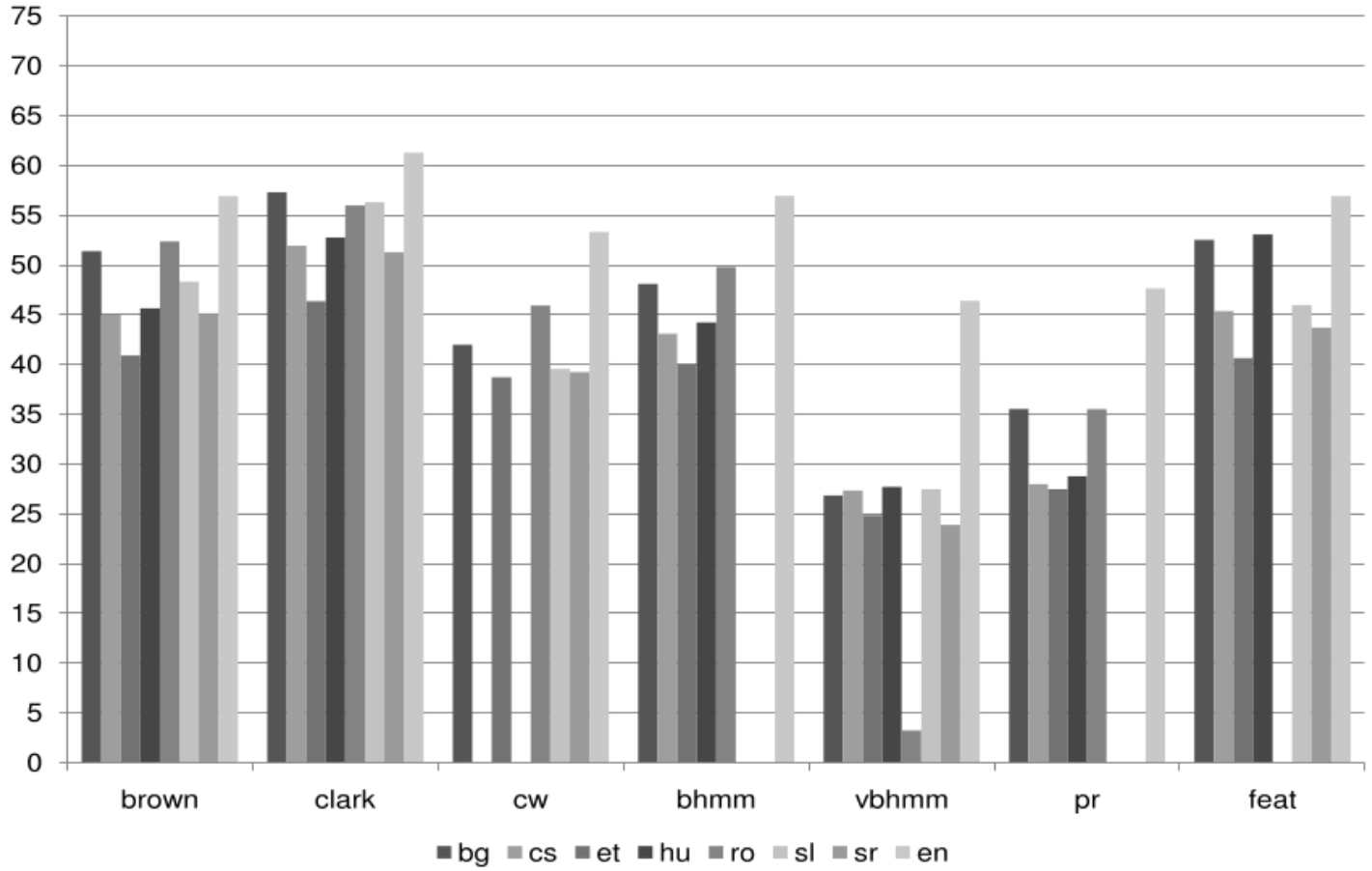


figure reproduced from Christodoulopoulos et al., 2010

Table of Contents

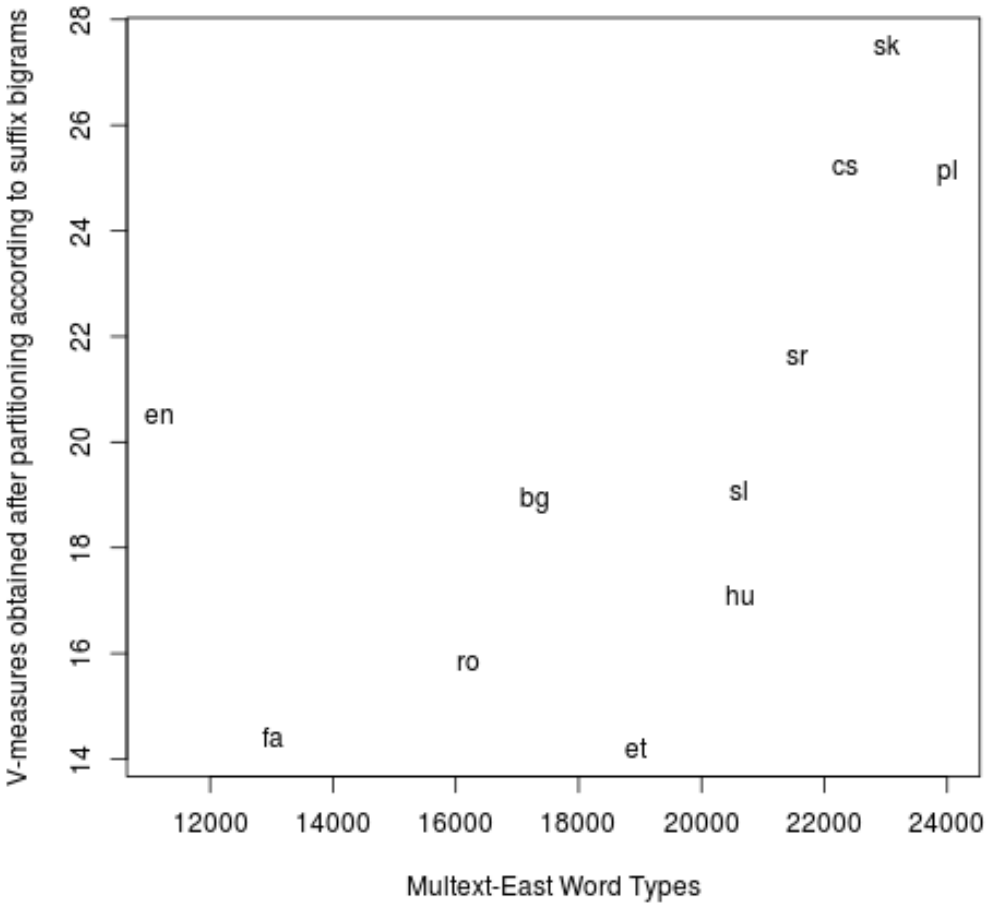
Advantages of suffix-started approach

- fast...very fast
- simple and therefore reliable
- can potentially "seed" more complex POS-i and GI systems
- multilingual and potentially universal
- consistent with operatory principle A

Operatory principle A

One of the first operating principles employed in the ontogenesis of grammar [is that] grammatical realizations in the form of suffixes or postpositions will be acquired earlier than realizations in the form of prefixes or prepositions (Slobin, 1973).

A correlation?



Possible correlation between number of word types a language has and usefulness of its suffixes for POS-i (Pearson co-efficient 0.597, p=0.0522)

Repetitio est mater studiorum

- surface features can be used for POS-i of diverse languages
- English is more an outlier than a centroid language of Multext-East corpus
- in all 11 studied languages, suffixes seem to yield useful cues for attribution of token into a POS-category and they are particularly useful for POS-i of Western Slavic languages
- according to certain psycholinguistic theories acquisition of suffixes is one of the first steps in ontogeny of grammar
- INDICATED CORRELATION: ? more word types the language has, higher the usefulness of suffixes in POS-partitioning ?

c.f. Hromada, Daniel Devatman. 2014. Conditions for cognitive plausibility of computational models of category induction. Proceedings of 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU2014). Springer.

Thank You for Your attention.

(I shall present here another paper tommorrow at 9:00)

slava Ukrajine!

 Berg-Kirkpatrick, Taylor, Alexandre Bouchard-Côté, John DeNero, et Dan Klein. 2010. Painless unsupervised learning with features. P. 582–590 in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

 Biemann, Chris. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. P. 7–12 in Proceedings of the 21st International Conference on computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop.

 Brown, Peter F., Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, et Jenifer C. Lai. 1992. Class-based n-gram models of natural language. P. 467–479 in Computational linguistics 18(4)

 Christodoulopoulos, Christos, Sharon Goldwater, et Mark Steedman. 2010. Two Decades of Unsupervised POS

induction: How far have we come?. P. 575–584 in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.


Clark, Alexander. 2003. Combining distributional and morphological information for part of speech induction. P. 59–66 in Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1.


Elman, Jeffrey L. 1989. Representation and structure in connectionist models.


Erjavec, Tomas. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. P. 131–142 in Language resources and evaluation 46(1)


Goldwater, Sharon, et Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. P. 744 in

Annual Meeting of Association of Computational Linguistics, vol. 45.


Graca, Joao, Kuzman Ganchev, Ben Taskar, et Fernando Pereira. 2009. Posterior vs. parameter sparsity in latent variable models. P. 664–672 in Advances in Neural Information Processing Systems 22.


Hromada, Daniel Devatman. 2014. Conditions for cognitive plausibility of computational models of category induction. Accepted for 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU2014). Montpellier, France.


Johnson, Mark. 2007. Why doesn't EM find good HMM POS-taggers. P. 296–305 in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

Karypis, George. 2002. CLUTO-a clustering toolkit.

Rosenberg, Andrew, et Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. P. 420 in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), vol. 410.

deSaussure, Ferdinand. 1922. Cours de linguistique générale. Payot, Paris.

Schütze, Hinrich. 1993. Part-of-speech induction from scratch. P.251–258 in Proceedings of the 31st annual meeting on Association for Computational Linguistics.

Slobin, Dan. 1973. Cognitive prerequisities for acquisition of grammar. P. 175-208 in Studies of child and language development.