Comparative study concerning the role of surface morphological features in the induction of part-of-speech categories

Daniel Devatman Hromada[1,2]

[1] Université Paris 8, Laboratoire Cognition Humaine et Artificielle, 2, rue de la Liberté 93526, St Denis Cedex 02, France
[2] Slovak University of Technology, Faculty of Electrical Engineering and Information Technology, Department of Robotics and Cybernetics, Ilkovičova 3, 812 19 Bratislava, Slovakia

Abstract. Being based on English language, existing systems of part-of-speech induction prioritize the contextual and distributional features "external" to the word and attribute somewhat secondary importance to features derived from word's "internal" morphologic and orthotactic regularities. Here we present some preliminary empirical results supporting the statement that simple "internal" features derived from frequencies of occurrences of character n-grams can substantially increase the V-measure of POS categories obtained by repeated bisection k-way clustering of tokens contained in Multext-East corpora. Obtained data indicate that information contained in suffix features can furnish c(l)ues strong enough to outperform some much more complex probabilist or HMM-based POS induction models , and that this can especially be the case for Western Slavic languages.

Keywords: part-of-speech induction, development of morphology, clustering, surface features, suffix

1 Introduction

Part-of-speech (POS) induction is a constructivist process aiming to converge to the mechanism able to attribute the POS category (e.g. "verb", "noun", "adjective" etc.) membership information to any word of the language under study. Because "syntactic category information is part of the basic knowledge about language that children must learn before they can acquire more complicated structures" [15] POS induction (POS-i) is often considered to be the first step in a more complex process of grammar induction and language acquisition in general.

Given such an important place of POS-i in NLP studies, it is of no surprise that while first computational models of POS-i were proposed decades ago [3][6][15] the problem of unsupervised POS-label attribution still attracts attention of many computational linguists. Thus, dozens of POS-i systems exist, among which those based on class-based word n-grams [5], graph clustering [2]

or diverse extensions to Hidden Markov Models [9][8][1] are compared in the [4] comparative study which suggests that "some of the oldest (and simplest) systems stand up surprisingly well against more recent approaches".

Aims of this article are 1) to elucidate a superior peformance of Clark [5] and Berg-Kirkpatrick [1] models with the statement: "Their models perform better because they use better features" 2) to precise that for many languages, such features can be morphological ones. We precise that what shall be called "morphological feature" (MF) in the rest of this article is any feature "internal" to the word WITHIN which it occurs and as such can be opposed to contextual or distributional features "external" to the word under study (i.e. opposed to features which describe word's relation to other words and not its internal composition).

By focusing upon the role of such "orthotactic" MFs in diverse languages represented in the Multext-East corpus [7] we shall try to persuade the reader that while the "syntax-in-word-order paradigm" could (and did) yield useful models and tools for description of English language, the uncritical acceptation of such paradigm could turn to be somewhat contra-productive if one tends to develop POS-i models for highly flectional & morphology-rich languages.

2 Corpus

All analyses were effectuated with texts contained in the 4th version of Multext-East corpus [7] . Bulgarian (bg), Czech (cs), English (en), Estonian (et), Farsi (fa), Hungarian (hu), Polish (pl), Romanian (ro), Serbian (sr), Slovak (sk) and Slovene (sl) transcription of Orwell's 1984 were analysed. Quantitative descriptions of different corpora are present in the table 1.

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
sk	23015	103452	13
sl	20597	112278	13
sr	21540	126611	13

3 Method

Every word from the corpus was described by a vector of features whose values were obtained by application of feature filters described below. Vectors were subsequently clustered into groups.

3.1 Feature extraction

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, ˆ and $ characters were prefixed, respectively suffixed, to extracted tokens. Following features were then extracted from tokens:

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

Character n-grams of length X [$\mathbf{N_x}$] – every feature encodes the number of occurrences of the character n-gram of length L within the token. Thus, if X=1, the word "ˆ 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 feature the number of "o" etc. If X=2, the vector could be [1, 1, 1, 1, 1], its first element representing the frequency of occurrence of "ˆ g" character bigram, second of "go" bigram, third of "oo" bigram etc.

Character fragments whose length <X [$\mathbf{F_X}$] – this approach takes into account all n-gram fragments BELOW the specified length X. Thus if X=3, the word "ˆgood$" could be represented by the vector [1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1] whose last four elements encode the presence of trigrams "ˆgo", "goo", "ood" and "od$"; composition of first 10 elements is explained above.

All fragments [A] – same as above but X is equal to word's length. Word's vector thus encodes occurrences of all 1gram, 2gram, 3gram ... X-gram character sequences present within the word. Yields biggest number of features.

Prefixes of length X [$\mathbf{P_X}$]– same as N_X but fragments of length X were extracted only from word's beginning

Suffixes of length L [$\mathbf{S_X}$]– same as P_X but fragments of length L were extracted only from word's beginning Word's circumference n-grams of length X [B_X] – boundary n-gram feature is a conjunction of a prefix and suffix feature, e.g. the B_2 feature for the word good can be matched by regular expression /ĝ.+d$/ and its occurrence would be also observed in the words like "god" or "gold"

Word's circumference [$\mathbf{C_X}$]– Conjunction of P_X and S_X,i.e. feature is defined by combination of prefix and suffix both of length X.

Word's root [$\mathbf{R_X}$]– for the purpose of this article, we define the root feature "as all that rests in the token when its circumference n-grams of length X are removed"

Token's co-occurrence neighborhood of length L [$\mathbf{O_L}$] – this is the only feature "external" to the token under study. Every co-occurrence of the definiens-token (column) maximum L words aside to the left or right from definiendum-token (row) augments the value by 1.

If the definiens does not co-occur aside the definiendum word or if a fragment (column) does not occur within the word, or a feature-representing pattern (column) does not match the word (row), then the value in the final vector is, of course, zero.

3.2 Clustering

Since our objective is to evaluate the (non)relevance of diverse sets of surface features for POS-i in different languages, and not to evaluate the subsequent grouping machinery, we have decided to use a simple (& fast) repeated bisection k-way clustering as is implemented in the clustered tool CLUTO [12]. 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.

4 Evaluation

For the purposes of this article we had decided to present our simulations principially in terms of V-measure. More theoretical [13] and empiric [4] reasons being explained elsewhere, our choice was partially motivated by the form of V-measure score equations:

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

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

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

which strongly resembles the F-measure score often used in evaluation of classification problems. The homogenity (h) and completeness (c) were designed in order to be analogic to precision, respectively recall. Given its elegance, stability in regards to growing number of clusters but also certain "strictness" (note that even the best state-of-the-art present in [4] comparative study rarely surpass the V>0.6 limit), we consider the Vmeasure to be very valuable quantitative measure of performance of clustering POS-i algorithms.

	L	**N_1**	**N_2**	**N_3**	**N_4**	**F_2**	**F_3**	**F_4**	**A**	**P_2**	**P_3**	**S_2**	**S_3**	**C_2**	**C_3**	**R_2**	**R_3**	**O_1**
bg	4.3	5.6	13.1	17.0	11.9	8.5	14.4	14.7	14.6	6.7	5.0	**18.9**	16.5	3.8	2.3	3.4	3.0	12.5
cs	5.4	9.2	**25.2**	20.7	11.6	23.1	24.8	23.9	24.3	7.4	7.1	**25.2**	18.7	4.7	3.1	3.7	3.4	7.9
en	3.8	6.5	14.1	15.3	9.4	10.4	14.9	16.1	14.7	3.9	3.6	**20.5**	19.7	2.4	1.7	2.9	2.2	14.4
et	4.2	4.0	12.2	14.2	11.9	5.8	6.92	9.38	7.24	4.2	6.0	14.2	**16.1**	3.6	2.8	3.4	3.3	6.77
fa	2.6	6.8	15.4	15.52	12.2	12.0	15.51	15.3	**15.55**	11.7	14.5	14.4	12.0	6.4	4.6	2.8	3.2	14.3
hu	2.3	4.3	6.1	10.7	9.4	5.2	6.26	6.58	5.65	5.4	5.7	**17.1**	14.2	3.0	1.8	2.4	2.0	7.1
pl	4.7	8.0	21.1	20.1	13.7	18.5	20.3	19.7	15.6	5.3	6.5	**25.1**	22.7	4.0	3.0	3.3	2.9	7.9
ro	4.6	7.1	11.1	13.6	9.5	8.23	11.3	11.8	10.9	6.5	5.9	**15.8**	14.8	3.1	1.9	2.5	2.4	15.6
sr	5.2	5.5	13.3	14.8	10.5	5.67	8.06	8.82	5.95	6.1	6.4	**19.1**	16.5	4.6	3.0	4.7	3.5	9.4
sk	5.9	11.2	26.9	21.0	14.0	23.8	24.9	24.2	22.5	8.2	5.8	**27.5**	21.3	4.8	3.5	3.6	3.5	8.7
sl	4.5	4.8	12.2	17.1	12.8	7.39	8.42	14.3	7.5	6.8	6.0	**21.6**	19.3	5.2	2.4	3.3	3.4	9.1

Table 1: V-measures obtained after clustering different corpus according to different features. The most performant feature of every corpus is marked.

Table above shows V-measure*100 values obtained by clustering of words characterized by length (L), character n-gram fragments of fixed (N_2, N_3, N_4) length or n-gram fragments shorter than certain length (F_2, F_3, F_4) as well as of clusters created by considering all fragments (A).

The best results (i.e. highest V-measures) were observed in case of Western Slavic languages which have all attained >0.2 of V-measure performance when clustered according to features representing character bigram occurrences. Southern Slavic languages along with Romanian, Hungarian and Estonian performed the best when character trigrams were taken into account. English attained the 0.16 performance when all bigrammata, trigrammata and tetragrammata were taken into account while Farsi was clustered the best when all n-gram character fragments were taken into account.

Further results presented in the table below point in the same direction. Highest V-measure score was attained by Slovak, Czech and Polish when simple extractor of suffix features of length 2 was applied. In fact the same extractor yielded highest scores in case of all languages with exception of Estonian where somewhat longer suffixes tend to facilitate the POS-i, and in case of Farsi whereby prefixal features seem to be at least as important as suffixal features. Word circumference features C_2 and C_3 as well their "negation", the word root features R_2 and R_3 do not seem to bring any information relevant to the categorization process – in fact they seem to perform even worse than the baseline feature L.

Members of set of "external" distributional features (O_1), which represent the trivial frequency of occurrence of the feature-word to the left or right from the target word, performed worse in all cases, English included, than S_2 .

5 Discussion

POS-i system comparative study of [4] indicates that POS-i models involving morphological features perform better than models which do not. However both in Clark's [5] probabilist model as well as in morphology-enriched HMM-derived [1] model, morphological features seem to play rather a role of a performance-increasing "cherry added to the top of the cake" than that of model's cornerstone.

Results presented in this paper suggest that focusing upon the phenomena occurring within the token, if the token's transcription allows it[3], seem to yield quite strong c(l)ues for subsequent clustering of tokens into their respective syntactic categories. It may be the case that especially the character bigrams occuring at word's offset position – suffixes – seem to play an important role in word→ POS category attribution. It is also worth noting that suffixes augment the performance of POS-i not only for Indo-European languages but also for Uralic languages like Estonian or Hungarian.

[3] For example, an "internal" feature-oriented approach would hardly yield any interesting results if applied on Chinese logograms but could be of certain theoretic interest when applied upon pinyin transcription.

It is also worth reiterating that POS-i within Western Slavic languages tends to be much more sensitive to character N-gram and suffix-derived features than other languages compared in this study. Because the research presented hereby was based only on one particular litteral corpus (Orwell's 1984) and the results obtained may thus represent not the properties of languages as such, but rather a certain translation style, it would be somewhat hors propos to postulate that a kind of overall statistic property - labeled hereby as "word offset flectivity" - is more marked in Western Slavic languages than, for example, in Southern Slavic or Uralic languages. But given the fact that it was only Slovak, Czech and Polish whose V>0.25 when clustered according to outputs of S_2 feature-extracting prism, we believe that subsequent analyses involving more corpora and more languages may be worth the effort. Verily only more exhaustive comparative studies could assess the impact of morphology of word X upon the attribution of syntactic function to the very word X. And since syntax is often bound with semantics – for example by means of thematic relations – such studies, if ever they would verify and not falsify the results presented hereby, could possibly result in a partial revision of a canonical "signifiant is independent from signifié" paradigm [14].

To emit such a call was, however, not a motivation behind the redaction of this paper. Nor had we aimed to outperform existing distributional&probabilist models – for it may seem quite unprobable that one would outperform the "heavy Markovian artillery" with such a simple computational machinery as k-way clustering. Thus, it has been of certain surprise to us that the comparison of data presented on Figure 4 in [4] with our results indicated that for some Slavic corpora, our simplistic morphology-driven geometrically-clustered model has attained higher or more or less equal V-mesure scores than models presented in [11][9]. Our approach can also dispose of certain advantages when it comes to computational complexity – while some models like that of [2] have sometimes problems to converge to result in reasonable time, none of our 198 analyses whose results are presented above have lasted more than few seconds on an average desktop computer.

This being said, we believe that it may be the case that POS-i induction of systems of next generation could not only take into account but shall rather be based on word's "internal" morpho(phono)logical or even prosodic and metric features. While sufficient evidence exists for stating that in order to have a highly performant and robust POS-i model, one MUST take into account the distributional and contextual information "external" to the word under question, we believe that especially in case of highly flectional languages, the complexity of the whole POS-i clustering proccess could be significantly reduced if ever the process shall be "seeded" (i.e. initiated) with token's "internal" features. Since the performance-augmenting and complexity-reducing effects of such seeding are the principal topic of our ongoing work, we conclude that what we believe to be the ultimate advantage of such a model could be its "cognitive plausibility" [10].

At last but not least, by underlining the importance of suffixal features for POS-induction process, our results may well point in the same direction as hy-

pothesis that "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"[16]. Thus, without an intention to do so[4] we ultimately find the results of our purely empiric study to be consistent with more general psycholinguistic theories of grammar induction and language development.

References

1. 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.
2. 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.
3. 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)
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.
5. 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.
6. Elman, Jeffrey L. 1989. Representation and structure in connectionist models.
7. Erjavec, Tomas. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. P. 131–142 in Language resources and evaluation 46(1)
8. 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.
9. 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.
10. 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.
11. 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).

[4] Both during conception and realization of our study, we have been utterly unaware neither of Slobin's "operating principle A", nor of amount of scientific evidence already associated with it.

12. Karypis, George. 2002. CLUTO-a clustering toolkit.
13. 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.
14. deSaussure, Ferdinand. 1922. Cours de linguistique générale. Payot, Paris.
15. Schütze, Hinrich. 1993. Part-of-speech induction from scratch. P.251–258 in Proceedings of the 31st annual meeting on Association for Computational Linguistics.
16. Slobin, Dan. 1973. Cognitive prerequisities for acquisition of grammar. P. 175-208 in Studies of child and language development.