New York Science Journal

Websites:
http://www.sciencepub.net/newyork
http://www.sciencepub.net

Emails:
newyorksci@gmail.com
editor@sciencepub.net

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

A Survey Of Pos Tagging For Languages Other Than English

Jawairya Bukhari[1], Sadia Basar[2], Muhammad Abid Ali[1]

[1]Department of Computer Science, University of Peshawar, KPK, 25000,Pakistan
[2]Department of Computer Science, Institute of Management Sciences, Peshawar, KPK, 25000, Pakistan
sadiaa.khancs@gmail.com

Abstract: In this paper, we first had a overall study of existing POS tagging on different languages other than English as because till now, most of the research done on POS tagging is for English. We observed that even though the research on POS tagging for English is done exhaustively, but there are other languages that are progressing and are improving. The goal of this paper is to highlight the idea that POS tagging can be performed on any language having odd features with little consideration related to their syntactic, semantic, and pragmatic and many other issues. Keeping the illustrated issues in mind once can develop a very efficient POS tagger for respective language.
[Adeagbo, AP; Ofoegbu, EO; Dada, TM; Adegboye, LA. **A Survey Of Pos Tagging For Languages Oth er Than English.** *N Y Sci J* 2021;14(8):53-55]. ISSN 1554-0200 (print); ISSN 2375-723X (online) http://www.sciencepub.net/newyork. 9. doi:10.7537/marsnys140821.09.

Keywords: Survey; Pos; Tagging; Language; English

I. INTRODUCTION

POS (Part Of Speech) tagging is usually performed on English as the language of preference. It is so as researcher wants to evaluate these proposed models with the existing models of English as lot of work has been done on English since 1962s. However work done on other languages cannot be denied. There is abundance of work done in other languages related to POS tagging. For other languages there is a scarceness of data sources and developing tools for language with limited sources is a challenge but necessary [1].The literature ranges from sophisticated studies for well known languages (for instance Germany) to those which are in initial stages of development (for instance Vietnam). The effort on less studied languages is following the same track of improvement in NLP as that for English. They are taking up the same methodology that is used in POS tagging for English [2]

Natural Languages may be classified into fixed word order language for instance English and free order languages for instance Sanskrit [3] for fixed order languages generative grammars like CFG and tree adjoining grammars are used for modeling sentimental structures [4]. These procedures do not work well with free order languages as quite large numbers of rules are involved to perform language processing [3].

For successful text analysis of any language a deep awareness of syntactic and semantic issues are necessitated [5] performing POS tagging for other languages the very first thing is to develop understanding of basic morphology of a language and how the firm characteristics of a language effects the NLP steps in text analysis. The accuracy of tagging model varies depending on the tag set used and field of ground reality data [5]. In tagging systems of unusual languages the number of tags varies from a dozen to several hundred depending on the specificity the information provided by the tag [6] .Estimation of the size of proposed tag set for particular collection of language is an issue of merit study [7].

For languages such as English word level POS tagging seems adequate because words usually match ups with the syntactically applicable POS tag classes. But for different other languages, this observation is scarce as the syntactic appropriate POS tag classes do not inevitably match up with the words. In many languages words are frequently produced by concatenating smaller parts, which acts as free morpho syntactic units, each one having its own POS. English is an inflectionally weak language, so troubles arise mainly in association with uncertainty at the word class e.g. in decisive whether “left” should be tagged as an adjective, a noun , or a verb. Taggers and tagged corpora were afterward developed also for morphologically richer languages, such as Czech and Slovene [8,9].

The majority of the taggers agree with the difficulty of determining the syntactic part of speech of an existence of a word in context, but they cannot be determining the collection of acceptable word without any such context. So for languages which are very context sensitive and are wealthy in morphology the lexicon which list probable tags have to be building very large in order to symbolize the

language acceptably [10]. Languages that are highly inflectional a compromise had to be made concerning the characteristic of the language that should be describing the new tag set [11]. Brill tagger has revealed good outcomes for English and there is confirmation that rule-based taggers can accomplish better results than stochastic ones in the English language (Samuelsson and Voutilainen, 1997). Furthermore, there are few winning attempts to teach the Brill tagger to languages other than English, such as German (Schneider and Volk, 1998), French (Chanod and Tapanainen, 1995), Italian (Basili et al., 1996) and Estonian (Schneider, 1997)

The paper is organized as follows. The next session is a brief overview of the POS tagging performed in many other odd languages. Section 3 reports the conclusion which briefs about the consideration that must be focused before performing a POS tagging on any language of interest.

II. A Brief Overview of POS tagging in Other Then Languages

The work on POS tagging can be cited for different language families and groups. No way the groups presented below are specific and nor comprehensive.

Indo European Languages

Many POS taggers are present for Spanish. The first SPOS Spanish Part of Speech Tagger) was developed in 1995 which uses rule based approach and was used as a module in Pagloss Knowledge based Machine Translation System. It consisted of 11 tags [12]. An Unsupervised learning approach is then used for POS tagging which uses Brills algorithm as that was developed for English [15]. Many lexis of Spanish act as different POS in different context so the need of automatic tagger system is very important [9] .other stochastic techniques are also used to gain accuracy in Spanish work.

Brill tagger has been used to perform POS tagging on Swedish corpus of 53444 tokens. The paradigm used was then enhanced by raising the size of the lexicon [10]. Nevertheless an efficient pos tagger that uses stochastic approach is also developed whose accuracy is about 97% for all words and 92% for unknown words [11].

Greek language has a rich structured tag set. The Greek tag set consist of 36 tags, 56 tags, 146 and the largest is 584. large size of 584 different tags is a problem. one more problem with the Greek language is that there are various different words forms which lead to large number of unknown words FBT (Feature based multi tired) has been used to tag such a highly infected language [12]. Transformed based Error learning approaches has been also used to resolve the ambiguity of POS tagging in Greek which leads the result up to 95% [13]

Agglutinative and Inflectional Languages

The agglutinative or inflective languages such as Turkish, Czech, Finnish, and Hungarian entail some obscurity in language dealing due to the more multipart morphology and relatively free word order in sentences when weigh against with languages like English. Many constraint based methods for morphological disambiguation in Turkish have been applied [14, 15]. A trigram-based statistical model has also been used in morphological disambiguation of Turkish text [15]. A current work has used a decision list induction algorithm called Greedy Prepend Algorithm (GPA) to learn morphological disambiguation rules for Turkish [15]. Work is done on Japanese Czech and Slovene with rule based, hybrid and pure stochastic approaches [2]

Semitic Language

The words that exist in Semitic text are made up of concatenation of words segments. Each one which match up to POS group. The Semitic words possibly ambiguous with view to their segmentation over and above the POS tags allocate to every word, So POS tagging is very watchful task to deal with such language. HMM which is a stochastic approach has been shown accuracy of 89% to 97% [15].Hebrew and Arabic are Semitic language. is described in Alder and Elhadad [14].Thai and Chinese, and obtained an accuracy of 94.3% for a Thai corpus and an accuracy of 91.4% for a Chinese corpus [15].

Other less studies and progressing Languages

Afrikaans the tag set varies up to 139. POS tagging is also performed African [12] Telugu POS tag set include more number of POS tag labels since Telugu is immensely inflective language [13].The three Telugu Pos taggers Rule-based POS tagger, Brill Tagger , Maximum Entropy POS taggers are developed with an accuracy of 98.016%, 92.146%, and 87.818 respectively. However another approach of voting algorithm is also used to get better the accuracy result for the tagging process. Telugu is an agglutinative language in which the words are formed by joining morphemes together [15].

Conclusion

The mass of literature on POS is for English. An inexperienced application of POS tagger developed for English in mind may not all the time work for other languages. For that reason oddity of language should be taken into report and essential frame work must be personalized to these languages

while developing POS tagger. The accuracy rate of POS tagging in English is about 96% to 97% and for the other languages comparable accuracy can be obtained provided that the distinctiveness of these languages then English are handled carefully. As far as less studied Languages are concerned, non-availability of lexical assets is a tailback for POS tagging. The morph syntactic tagging of agglutinative or inflective languages is more complex due to the large number of tags.

The use of morphological features is especially supportive to develop a rational POS tagger when tagged resources are inadequate Those languages which have inadequate POS tagged corpora or limited recourses unsupervised POS tagging is an appropriate option, however hybrid approaches can also be used. Many languages code added information than just part-of speech tag in a word thanks to the more complex morphology. To deal with morphological disambiguation; we need to determine all the syntactic morphological features of a word. Therefore morphological disambiguation can be called morph syntactic tagging in analogy to part-of-speech tagging.

References

[1] Mukund, S. Srihani, R. "NE Tagging For Urdu Based On BootStrap POS learning". In Proceeding Of CUAWS3, Third International Cross Lingual Information Access Workshop, Boulder, Colorodo, June 2009. Pages;61-69.

[2] Ray, P.R., V,H., Sarkar, S., Basu, A. " *Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Processing"*. ICON 2003.

[3] Charniak, E. "Statistical parsing with a context-free grammar and word statistics". Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97), AAAI Press/MIT Press, Menlo Park. 1997

[4] John T. Platts A grammar of the Hindustani or Urdu language. Munistirian Delhi 1967

[5] Erjavec, T., Sárossy, T."Morphosyntactic Tagging of Slovene Legal Language". Informatica (Slovenia) 30(4): 483-488 (2006)

[6] Using Multiattribute Prediction Suffix Graphs for Spanish Part-of-Speech Tagging. In Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis. Pages 228-237 Springer-Verlag London 2001.

[7] Papageorgiou, H., Prokopidis, P., Giouli, V., & Piperidis, S. (2000). A unified POS tagging architecture and its application to Greek. In *Proceedings of the language and resources evaluation conference*.

[8] Aone,C., Hausaman,. K."Unsupervised learning of a rule-based Spanish Part of Speech tagger" Proceedings of the 16th conference on Computational linguistics - Volume 1. Pages 53-58.

[9] *Prütz, K.* Part-of-speech tagging for Swedish.

[10] Lu, B.L., Ma, Q.,Isahara,H., and Ichikawa M. "Efficient Part-of-Speech Tagging with a Min-Max Modular Neural-Network Model, in Applied Inteligence, 2001

[11] Oflazer, K., T¨ur, G.: Combining Hand-crafted Rules and Unsupervised Learningin Constraint-based Morphological Disambiguation. Proceedings of the ACLSIGDAT Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, USA (1996).

[12] Oflazer, K., T¨ur, G.: Morphological Disambiguation by Voting Constraints. Proceedings of ACL/EACL, The 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain (1997)

[13]. Hakkani-T¨ur, D. Z., Oflazer, K., T¨ur, G.: Statistical Morphological Disambiguation for Agglutinative Languages. Computers and the Humanities 36(4) (2002)

[14] Y¨uret, D., T¨ure, F.: Learning Morphological Disambiguation Rules for Turkish. Proceedings of HLT-NAACL (2006).

[15]Meni Adler and Michael Elhadad. 2006. An unsupervised morpheme-based HMM for Hebrew morphological disambiguation. In Proceeding of COLING-ACL-06, Sydney, Australia.

2/3/2021