FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 4 No. 2, June, 2020, pp 712 – 721
DOI: https://doi.org/10.33003/fjs-2020-0402-325

PARTS OF SPEECH TAGGING: A REVIEW OF TECHNIQUES

[*1]Jamilu Awwalu, [2]Saleh El-Yakub Abdullahi, [1]Abraham Eseoghene Evwiekpaefe

[1] Department of Computer Science, Nigerian Defence Academy, Kaduna State, Nigeria
[2] Department of Computer Science, Nile University of Nigeria, Abuja, Nigeria

[*]Corresponding Author's Email: awachi.jami@nda.edu.ng

ABSTRACT
Technology advances by the day and computers can be considered as valuable to almost every learned person. One of the most uses of computers nowadays is for internet surfing and social networking. Computers in this context are not restricted to desktop or laptop computers only. Internet surfing and social networking has made interactions between people and computers very easy, where people can communicate using their languages thus making processing of these languages a useful task for the computers to interpret. The correct processing of these languages on the computer relies on the correct identification of parts of speech (POS) in sentences which has been an active area of research for a long time. This paper presents a review of the different techniques used in parts of speech tagging that range from Unilingual to Multilingual Parts of Speech (POS) tagging approaches. The purpose of this study is to elaborate and compare the different tagging techniques in terms of their characteristics, difficulties, and limitation.

Keywords: Rule Based POS Tagging, Stochastic POS Tagging, Hybrid POS Tagging, Word Alignment, Code Switching.

INTRODUCTION

Parts of Speech Tagging (POS) according to Robin (2009) is the process of allocating a particular part of speech to a word. POS tagging works by assigning parts of speech label to words given in a text (Pandian and Geetha, 2008). The POS tagger according Manning et al. (2014) consist of three elements which are; annotation, assigning correct grammatical position for each word, and parsing. While some researchers build new NLP resources for the first time like Dione et al. (2010), others continued working to improve the available NLP resources for some languages like Albared et al. (2009). Dione et al. (2010) designed and implemented resources for POS tagging of Wolof, a language spoken in Senegal. This involved building the first publicly available NLP resources for the language which include Tagset and POS annotated gold standard from scratch.

TECHNIQUES OF POS TAGGING

Parts of Speech (POS) tagging have been implemented by several researchers using different techniques. Techniques such as Bayesian Models, Markov Models, Maximum Entropy, and Transformation-Based Learning (TBL) have been applied in POS tagging. Nguyen et al., (2016) applied TBL for POS tagging and achieved competitive accuracy values compared to the other techniques mentioned. While researchers such as Rathod and Govilkar (2015) and Kumawat and Jain (2015) grouped POS tagging techniques based on supervised or unsupervised technique as shown in figure 1, others such as Amri et al. (2017) broadly grouped POS techniques into five categories which are; Statistical approach, Rule Based approach, Hybrid approach, Transformation-Based Learning approach, and Memory Based approach. The Memory Based Approach as described by Khemakhem et al. (2016) assumes that words which occur in similar contexts will be assigned the same tag. It is a similarity-based supervised learning which is an extension and adaptation of classical k-Nearest Neighbor (k-NN). Mahar and Memon (2010) described supervised POS taggers as taggers built on pre-tagged corpora while the unsupervised POS taggers do not require any pre-tagged corpora, rather they employ methods that automatically tags assigned words. According to Das and Petrov (2011) supervised tagging rely on training data that is labeled, and this labelling of data consumes time and costs a lot to generate, on the other hand, the unsupervised approach to learning seem to be the likely solution to this problem of cost and time consumption. This is because the unsupervised approach only requires unannotated text for training models. Unsupervised POS taggers employ advanced computational methods such as Baum-Welch algorithms to induce tags automatically (Kumawat and Jain, 2015). However, the practical usability of Unsupervised POS taggers is questionable because the best English POS tagger that is completely unsupervised according to Das and Petrov (2011) achieved the limited accuracy of 76% as stated by Christodoulopoulos and Steedman (2010).

According to Khemakhem et al. (2016) the early approaches used for POS tagging are rule-based. Based on an architecture that is two-staged, the first stage uses the dictionary approach in assigning potential POS to each given word. The second stage uses hand-written list that contain rules for disambiguation that are used in arranging the given list to a part-of-speech for each word. The set of rules in this approach must be properly written and inspected by human experts. After the 80s, the Statistical (stochastic) approach came into existence and gained more popularity because it requires lesser work and it is not as costly as the rule-based approach.

Generally, the process of tagging according to Sonai et al. (2017) can be grouped into stochastic or rule based tagging. This is supported by figure 1 where Rathod and Govilkar (2015) identified rule based, stochastic, and hybrid as the techniques under both supervised and unsupervised approach to POS tagging.

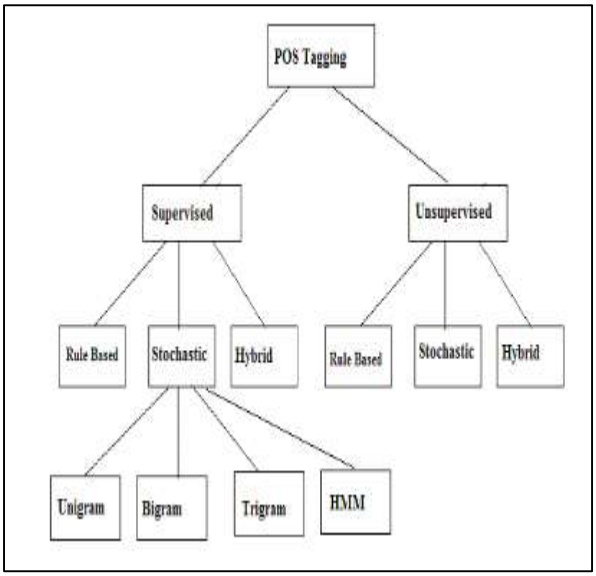

Figure 1 Classification of POS Tagging techniques. Source: Rathod and Govilkar (2015)

Rule Based POS Tagging

This approach according to Kumawat and Jain (2015) is the oldest in part-of-speech tagging system. The rule based approach consists of hand written set of rules that are used in assigning POS tags to words. These rules can include punctuation and capitalization patterns in a set of sentence or words. This approach as shown in figure 2 was implemented on Malay language POS tagging by Sonai et al. (2017).

Benefits of Rule Based POS Tagging

Some of the benefits of rule based tagging as stated by Brill (1992) are:

i. It significantly reduces the amount of information storage because it represents acquired knowledge in the form of rules, not stored data records. This means that a rule based model does not require a large tagged corpus in order to estimate probabilities values.
ii. A perspective that describes linguistic phenomena in an explicit way is used in writing the language model.
iii. The language model can comprise of several complex forms of knowledge.
iv. It is easier to understand and maintain written rules.
v. The rule based approach is highly portable from a text corpus to another.
vi. It allows the building of a highly accurate language system.

Limitations of Rule Based POS Tagging

Limitations of the rule based approach to POS tagging includes:

i. The necessity of linguistic background, and manual construction of rules (Kumawat and Jain, 2015). This makes it hard for researchers because of the time it takes to manually construct the rules.
ii. It cannot be guaranteed that every linguistic rule is captured in the rule construction.
iii. Language model transporting to other languages is less applicable (Brill, 1992)
iv. A high labour in terms of cost and work is usually required (Brill, 1992)
v. Information frequency is mostly not considered by the language models (Brill, 1992)

Stochastic POS Tagging

Stochastic POS tagging is implemented based on different models such as Hidden Markov Model (HMM), Maximum Entropy, Maximum Likelihood Estimation, Support Vector Machines, Conditional Random Fields, and N-grams, this technique of POS tagging makes use of statistics, frequency, and probability (Kumawat and Jain, 2015). An approach to stochastic POS tagging implemented on Malay language is shown in figure 3

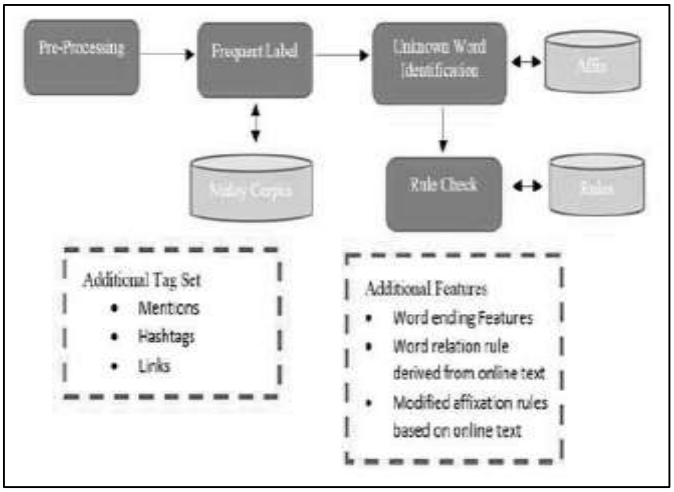

Figure 2 Rule Based POS Tagging (Sonai et al., 2017)

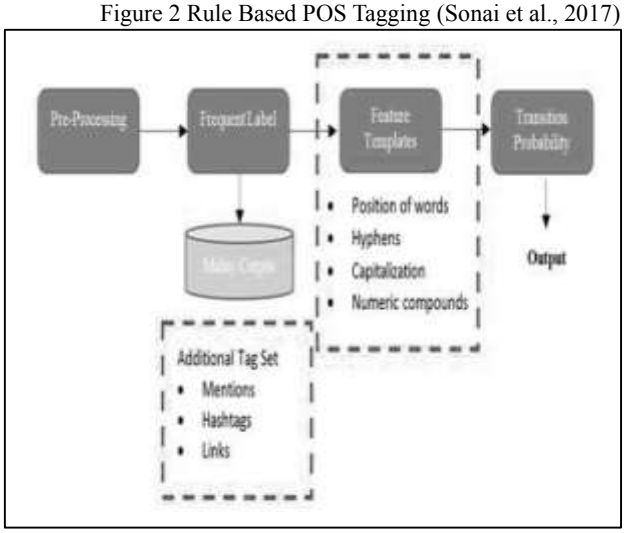

Figure 3 Stochastic POS Tagging (Sonai et al., 2017)

A study by Dalal et al. (2007) presented a statistical POS tagger for Hindi language using Maximum Entropy Markov (MEM) with a rich set of features. The accuracy of the statistical model was 94.89%, making it a score that is high and at that time the highest reported accuracy for Hindi POS taggers. Another research carried out by Md. and Hasan (2014) on stochastic POS tagging of techniques used in Bengali, which is a language with a very high number of speakers around the world. In the work, they reviewed types of corpus and tags used and their methods out of which they found a hybrid Hidden Markov Model (HMM) with morphological analyser works best for Bengali with accuracy of 96.3%.

Benefits of Stochastic POS Tagging

Some benefits of stochastic tagging as stated by Merialdo (1994) are:

i. The transportation of a language model to another is easier when there is large manually tagged corpus available.
ii. Frequency information is considered by language models.
iii. Based on available data, probabilities of model can be automatically estimated.

Limitations Stochastic POS Tagging

The first three limitations presented below as stated by Merialdo (1994) are some of limitations of the stochastic approach:

i. Stochastic model require a large corpus that is manually tagged in order to calculate probabilities.
ii. The model requires a huge matrix in order to represent information.
iii. Cannot deal with unknown words.
iv. Incorrect sequence tags as per language grammar rules (Kumawat and Jain, 2015).

Hybrid POS Tagging

The Hybrid technique to POS tagging uses combination of two or more techniques to implement the tagging process. Different researchers have implemented various approaches to the hybrid POS tagging. Some researchers such as Dandapat et al. (2004) implement the fusion of supervised and unsupervised Hidden Markov Model (HMM) which is a stochastic POS tagger as a hybrid tagger for Bengali language while others such as Rathod and Govilkar (2015) and Btoush et al. (2016) view the hybrid approach as combination of rule based and stochastic techniques POS tagging. The Transformation-based as described by Btoush et al. (2016) is a type of hybrid POS tagger that combines rule-based and stochastic approach in POS tagging.

Previous studies recorded accuracies in employing the Markov Model (MM) as part of model they implemented, however on comparing the performance of the HMM with the Brill's tagger also known as the TBL, Hasan et al. (2007) worked on 4048 tokens where they found the Brill's tagger performed better than the HMM and the Tagger by achieving a higher accuracy. In a related work by Delic et al. (2009) on POS tagging of Serbian language, TBL achieved 10.00% error rate which they mentioned is comparable with accuracy recorded for similar research on other languages. Also, similar competitive accuracy was recorded by Nguyen et al. (2016) where they implemented TBL for 13 languages, further more achieving faster training and tagging speed.

Although no limitations to hybrid tagging approach have been clearly stated, some of the generic limitations of hybrid algorithms that cut across area of applications are:

i. Inherited limitations of specific methods or algorithms that are not resolved in the hybridization continue to exist in the hybrid algorithm.
ii. Increase in computation as a result of dependents of the combination of two or more methods or algorithms that are involved.
iii.

POS Tagging Performance Measure

There are difficulties associated with POS tagging varying from model accuracy to speed. In terms of accuracy, the two main issues that affect accuracy of a POS tagger are ambiguity and unknown words (Sonai et al., 2017). The performance measure of a POS Tagger as stated by Hladka (2000) is determined by tagging accuracy (TA) or error rate (ER). TA is the percentage of the correctly tagged words in a sentence, while ER is the percentage of incorrectly tagged words in a sentence.

COMPARISON BETWEEN POS TAGGING TECHNIQUES

Several comparative studies have been conducted on different POS tagging techniques, different corpus with varying token sizes, and different languages.

A study was conducted on Bangla text by Kumawat and Jain (2015) on comparing N-gram which is rule-based, Hidden Markov Model (HMM) which is a stochastic approach, and Brill's tagger which is hybrid approach. Results from the work shows the N-gram model recording lesser accuracy compared to HMM in their experiments with performance increasing with the increase of corpus size. However, the Brill tagger outperforms both N-gram and HMM models in all their experiments. Contrary to the findings of Kumawat and Jain (2015), a study by Sonai et al. (2017) recorded better accuracy on rule based POS taggers in comparison to stochastic POS taggers. The study which was on Malay text by Sonai et al. (2017) to find out performance of the two techniques on ambiguous and unknown Malay online text from 500 tweets containing 5850 words out of which 6.80% (397) are unknown words. Both stochastic and rule based taggers achieved same accuracy on known words, but on unknown words; the rule based tagger achieved 5% and 2.1% more than stochastic tagger on unknown words and ambiguous words respectively.

A study by Hasan et al. (2007) compared different techniques on Bangla language similarly supports results from Kumawat and Jain (2015). As stated by Hasan et al. (2007), results on the Bangla corpus from their study was not satisfactory because it was small sized. As the results of their study are similar to that of Kumawat and Jain (2015), the two study further reveals more similarity, which is; increasing accuracy as corpus size increases. The results from Hasan et al. (2007) revealed that the Rule based approach outperformed the HMM Stochastic approach. This was even as the HMM algorithm was trained on the Brown corpus which is a very large English language corpus.

A study by Rathod and Govilkar (2015) on Indian languages as shown in table 1 revealed comparative differences based on techniques employed by previous studies of different researchers.

Table 1 Comparison of POS Tagging Techniques on Indian Languages (Rathod and Govilkar, 2015)

Technique	Description	Advantages	Disadvantages	Accuracy
Rule Based	Uses a set of hand written rules.	i. Small set of simple rules ii. Less stored information	Generally less Accurate as Compared to Stochastic taggers.	i. Marathi - 78.82% ii. Sindhi - 96.28% iii. Sanskrit – 90.00%
Stochastic	Probabilistic Depending on the number of previous tags (1, 2,and 3) Called , bigram or trigram frequencies in a training corpus.	i. Generally more accurate as compared to rule based taggers	i. Relatively complex. ii. Require vast amounts of stored information	i. Marathi, Bigram, Trigram and HMM gives the accuracy of 77.38%, 90.30%, 91.46% and 93.82% respectively. ii. Bengali bigram tagger-74.33 and trigram - 78.68
Hybrid	Assign the most probable tag to the word using statistical after that, if wrong tag is found then by applying some rules tagger tries to change it.	i. Having higher Accuracy than individual rule based or statistical approach	i. Not assign correct tag to an unknown word	i. Hindi - 79.66% ii. Bengali – 95.00%

MULTILINGUAL POS TAGGERS

The key hypothesis of multilingual learning as stated by Snyder et al. (2008) is that "by combining cues from multiple languages, the structure of each becomes more apparent". Several works have been conducted on multilingual POS tagging ranging from unilingual, bilingual to POS taggers that accommodate three or more languages. A research by Naseem et al. (2009) employed unsupervised POS tagging on multilingual approach in order to make apparent the inherent patterns of ambiguity in assigning parts of speech tags. Two models; (a) merging tag structures for language pairs into a single pair by employing joint distributions over aligned tag-pairs and (b) using latent variables instead of explicit node merging to incorporate multilingual contexts, that were formulated as hierarchical Bayesian models using Markov Chain Monte Carlo sampling inference technique. Evaluating the models on eight languages Bulgarian, Czech, English, Estonian, Hungarian, Romanian, Serbian, and Slovene, results show impressive achievement from incorporating the multilingual approach. Also steady increase in performance as number of available languages increases.

WORD ALIGNMENT AND CODE SWITCHING IN MULTILINGUAL POS TAGGING

MADAMIRA was proposed by Pasha et al. (2014) as a supervised morphological disambiguator or a POS tagger for the Arabic language text. It extracts a set of different associated linguistic and morphological information from given input. This includes part-of-speech information, detailed morphology, lemmas, phrase-level information such as base phrase chunks, fully-diacritized forms, and named entity tags (Alghamdi et al., 2016). Although MADAMIRA produces a rich output, it has the limitation of being slow (Khalifa et al., 2016).

A study by Alghamdi et al. (2016) implemented POS tagging for two pairs of Code Switched data; they are Spanish – English, and Modern Standard Arabic (MSA) – Arabic dialects. In the study they used the TreeTagger to train Penn Treebank data for English, and Ancora-ES for Spanish. MADAMIRA was used to compose the Arabic dataset. Publicly available in two versions; MSA and EGY, the version of MADAMIRA MSA was trained on newswire data (Penn Arabic Treebanks) (Maamouri et al., 2004) while MADAMIRA EGY is trained on Egyptian blog data which comprises a mix of MSA, EGY and CS data (MSA-EGY) from the LDC Egyptian Treebank parts 1-5 (ARZ1-5) (Maamouri et al., 2006). Results from the study reveals that depending on the language pair and the distance between them, there are varying degrees of need for annotated code switched data in the training phase of the process and that when code switched, languages that share a major amount of homographs will benefit from more code switched data at training time, while languages that are distant apart such as English and Spanish, when codeswitched, benefit more from having larger monolingual data mixed.

To overcome the slow performance of MADAMIRA, different systems were proposed by different researchers. Abdelali et al. (2016) proposed Fast and Furious Segmenter for Arabic (FARASA) based on SVM-ranks using linear kernels. Trained and tested on English – Arabic dataset from the Penn Arabic Treebank (ATB), FARASA was compared with MADAMIRA and Stanford Arabic segmentor on Machine Translation (MT) and Information Retrieval (IR). Results show that FARASA is at least an order of magnitude faster than both MADAMIRA and Stanford Arabic segmentor, and it also performs slightly better in intrinsic evaluation.

Similarly, a study by Khalifa et al. (2016) proposed Yet Another Multi-Dialect Arabic Morphological Analyze (YAMAMA) based on same datasets used for MADAMIRA as an alternative to FARASA and MADAMIRA. Result from the proposed system show that YAMAMA is almost five times faster than the state-of-the-art MADAMIRA system but with a slightly lower quality. YAMAMA uses a pre-computed maximum likelihood model to as-sign an analysis to every word. For out-of-vocabulary words, YAMAMA ranks all of the analyses for such words using two language models of the lemma and the Buckwalter POS tag.

Word alignment and Graph projection are used in bilingual and multilingual NLP tasks. A study by Bigvand et al. (2017) employed word alignment in implementing semi-supervised model that adds alignment type information to word alignments using two datasets i.e. GALE Chinese – English and Hong Kong parliament datasets (HK Hansards). On word alignment for POS tagging, a corpus based approach by Dien and Kiem (2003) on English – Vietnamese POS tagging used TBL by employing English – Vietnamese parallel corpus (EVC) as shown in figure 4 to bootstrap the POS-annotation results of the English POS-tagger by exploiting the POS-information of the corresponding Vietnamese words via their word-alignments in EVC. The process flow of the study is shown in figure 5. Results from applying the model on 1000 manually annotated words recorded 94.6% accuracy.

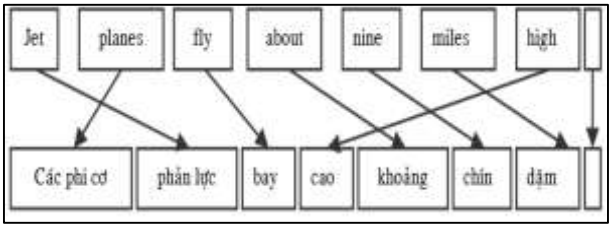

Figure 4 Word-aligned pair of sentences in EVC corpus, Dien and Kiem (2003)

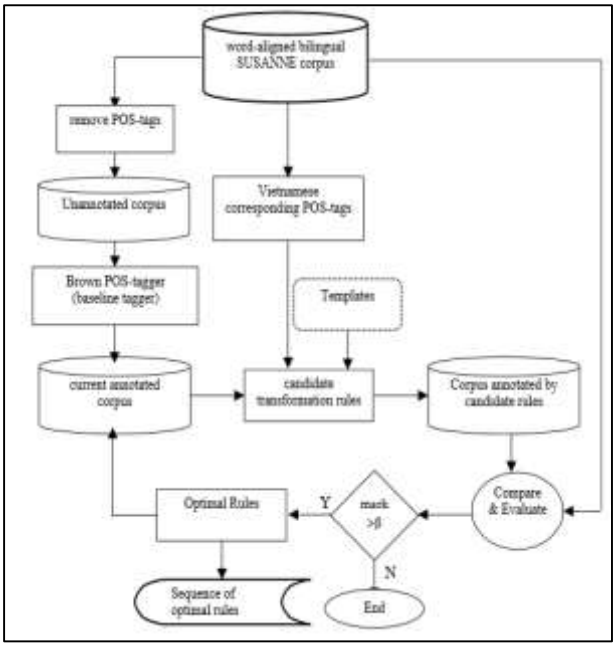

Figure 5 Flowchart of TBL-algorithm in POS-tagger for EVC corpus.
Source: Dien and Kiem (2003)

Similarly, Khemakhem et al. (2016) used aligned corpora approach for knowledge transfer in implementing POS tagging for under-resourced languages. The experimentation of their proposed approach was performed for the language pair: Arabic as an under-resourced language and English as a rich-resourced language. The approach of their study as shown in figure 6 and 7, annotate the source side of the aligned parallel corpus, then generate the tags for the target side by using word alignments by employing the corpus based method also known as the Memory Based Learning (MBL).

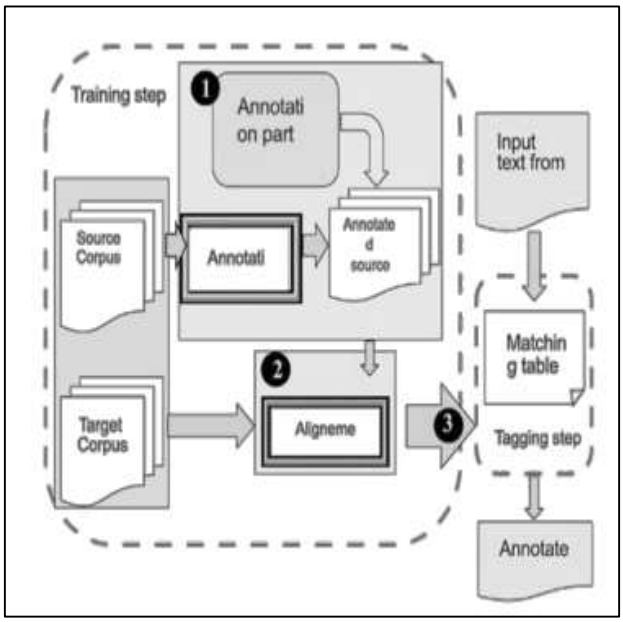

Figure 2.6 Proposed Approach to POS Tagging.
Source: Khemakhem et al. (2016)

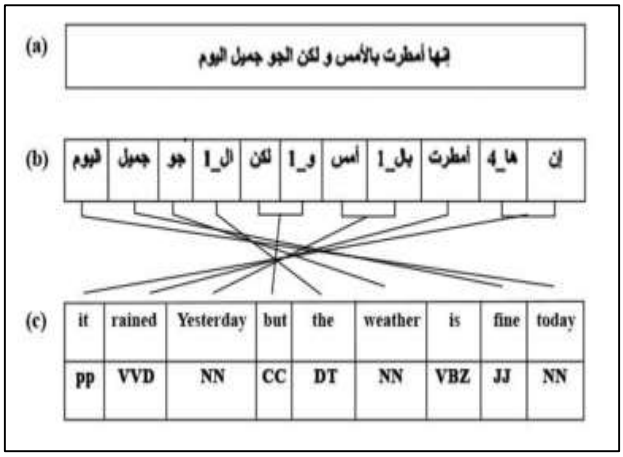

Figure 2.7 Word alignments example by Khemakhem et al. (2016), (a) Original Arabic sentence, (b) segmented Arabic sentence, (c) English translation and its alignment with a morphological analysis

A major challenge of POS tagging Code Switching (CS) data at inter or intra sentential level as described by Çetinoglu et al. (2016) is the lack of large annotated data. One approach that proved to be quite useful in previous works according to Çetinoglu et al. (2016) is the use of Language IDs (LIDs). Soto and Hirschberg (2018) modified the input space of Bi-directional Long Short-Term Memory (Bi-LSTM) tagger to make use of the language ID information to tag POS for switched data of English-Spanish Languages based on Recurrent Neural Network (RNN). The model was trained on two multilingual corpora i.e. Universal Dependencies (UD) Corpora, Miami Bangor (MB) and one monolingual corpus i.e. Wall Street Journal (WSJ). Results from the study show that (a) the monolingual taggers trained on benchmark training sets perform poorly on the test set of the CS corpus, (b) CS models achieve high POS accuracy when trained and tested on CS sentences, (c) expanding the feature set to include language ID (LID) as input features yielded the best performing models, (d) a joint POS and language ID tagger performs comparably to the POS tagger and its LID accuracy is higher than 98%, and (e) a model trained on instances of in-genre inter-sentential CS performs much better than the monolingual baselines, but yielded worse test results than the model trained on instances of inter-sentential and intra-sentential code-switching.

A study by Das and Petrov (2011) proposed using bilingual approach using Graph-Based projections. This approach to building NLP tools for resource-poor languages leverage on existing resource for a resource rich language to reduce the gap between accuracy of supervised and unsupervised tagging approaches as identified by (Christodoulopoulos and Steedman, 2010). The two sets of datasets were used in the study were (a) Monolingual treebanks, and (b) large amount of texts containing English on one side i.e. Danish, Dutch, German, Greek, Italian, Portuguese, Spanish and Swedish. Results from across eight European languages shows that the proposed approach results in an average absolute improvement of 10.4% over an advanced baseline, and 16.7% over hidden Markov models (HMM) induced with the Expectation Maximization algorithm.

CONCLUSION

POS tagging is an integral part of Natural Language Processing. A lot of work has been done on it for different languages using different approaches. In this review, we highlight techniques for single language POS tagging, Multilingual Parts of Speech (POS) tagging approaches, rule based, stochastic, and hybrid tagging techniques. Also, the comparison of the different tagging techniques, their characteristics, difficulties, limitation was presented. The different techniques discussed in this review have their limitations; as such there are no clear ways to implementing a technique without any limitations. This is because the limitations are more aligned to the algorithms used under the different tagging techniques. Therefore, overcoming these limitations is majorly a trade-off between the limitations avoided and those to be encountered. In general, this review would help researchers in comparing and deciding a tagging approach with minimal number of limitations as they aim to achieve their research objectives.

REFERENCES

Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa : A Fast and Furious Segmenter for Arabic. In *Proceedings of NAACL-HLT 2016 (Demonstrations)* (Vol. 2016, pp. 11–16).

Alghamdi, F., Molina, G., Diab, M., Solorio, T., Hawwari, A., Soto, V., & Hirschberg, J. (2016). Part of Speech Tagging for Code Switched Data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching* (pp. 98–107).

Amri, S., Zenkouar, L., & Outahajala, M. (2017). A Comparison of Three Machine Learning Methods for Amazigh POS Tagging, 83–87.

Bigvand Mansouri, A., Bu, T., & Sarkar, A. (2017). Joint Prediction of Word Alignment with Alignment Types. *Transactions of the Association for Computational Linguistics*, *5*, 501–514.

Brill, E. (1992). A Simple Rule-Based Part of Speech Tagger. *ANLP*, 152–155.

Btoush, M. H., Alarabeyyat, A., & Olab, I. (2016). Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *7*(6), 331–335.

Çetino, Ö., Schulz, S., & Vu, T. N. (2016). *Challenges of Computational Processing of Code-Switching*.

Christodoulopoulos, C., & Steedman, M. (2010). Two Decades of Unsupervised POS induction : How far have we come ? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 575–584).

Dalal, A., Kumar, N., Uma, S., Sandeep, S., & Pushpak, B. (2007). Building Feature Rich POS Tagger for Morphologically Rich Languages : Experiences in Hindi. In *5th International Conference on Natural Language Processing* (p. 9).

Dandapat, S., Sarkar, S., & Basu, A. (2004). A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali. *Transactions on Engineering, Computing, and Technology*, *VI*(December), 169–172.

Das, D., & Petrov, S. (2011). Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 600–609).

Dien, D., & Kiem, H. (2003). POS-Tagger for English-Vietnamese Bilingual Corpus. In *HLT-NAACL-PARALLEL '03 Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond* (pp. 88–95).

Hasan, F. M., UzZaman, N., & Khan, M. (2007). Comparison of different POS Tagging Techniques (n-gram, HMM and Brill's tagger) for Bangla. In *Advances and Innovations in Systems, Computing Sciences and Software Engineering* (pp. 121–126). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-1-4020-6264-3_23

Hladka, B. (2000). *Czech Language Tagging*. Institute of Formal and Applied Linguistics, Charles University.

Khalifa, S., Zalmout, N., & Habash, N. (2016). YAMAMA : Yet Another Multi-Dialect Arabic Morphological Analyzer. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System*

Demonstrations (pp. 223–227).

Khemakhem, I. T., Jamoussi, S., & Hamadou, A. Ben. (2016). POS Tagging without a Tagger : Using Aligned Corpora for Transferring Knowledge to Under-Resourced Languages POS Tagging without a Tagger : Using Aligned Corpora for Transferring Knowledge to Under-Resourced Languages. *Computacion Y Sistemas*, *20*(4), 667–679. http://doi.org/10.13053/cys-20-4-2430

Kumawat, D., & Jain, V. (2015). POS Tagging Approaches: A Comparison. *International Journal of Computer Applications*, *118*(6), 975–8887. Retrieved from http://research.ijcaonline.org/volume118/number6/pxc3903148.pdf

Maamouri, M., Bies, A., Buckwalter, T., Diab, M. T., Habash, N., Rambow, O., & Tabessi, D. (2006). Developing and Using a Pilot Dialectal Arabic Treebank. In *LREC*. inproceedings.

Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004). The penn arabic treebank : Building a large-scale annotated arabic corpus The Penn Arabic Treebank : Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR conference on Arabic language resourcesand tools* (pp. 466–467).

Mahar, J. A., & Memon, G. Q. (2010). Rule Based Part of Speech Tagging of Sindhi Language. In *Proceedings of the 2010 International Conference on Signal Acquisition and Processing* (pp. 101–106). inproceedings, Washington, DC, USA: IEEE Computer Society. http://doi.org/10.1109/ICSAP.2010.27

Manning, C. D., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60).

Md., A. kalaam, & Hasan, R. (2014). Review of Stochastic POS tagging techniques used in Bengali. *International Journal of Computer Applications*, *102*(8), 35–39.

Merialdo, B. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics*, *20*(2), 155–171.

Naseem, N., Snyder, B., Eisenstein, J., & Barzilay, R. (2009). Multilingual Part-of-Speech Tagging : Two Unsupervised Approaches. *Journal of Artificial Intelligence Research*, *36*, 341–385.

Pandian, S. L., & Geetha, T. V. (2008). Morpheme based Language Model for Tamil Part-of-Speech Tagging. *Polibits*, *38*, 19–25.

Pasha, A., Al-badrashiny, M., Diab, M., Kholy, A. El, Eskander, R., Habash, N., … Roth, R. M. (2014). MADAMIRA : A Fast , Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of LREC, Reykjavik, Iceland.* (pp. 1094–1101).

Rathod, S., & Govilkar, S. (2015). Survey of various POS tagging techniques for Indian regional languages. *International Journal of Computer Science and Information Technologies*, *6*(3), 2525–2529.

Robin. (2009). World of Computing. Articles on Natural language Processing. Retrieved February 28, 2018, from http://language.worldofcomputing.net/pos-tagging/parts-of-speech-tagging.html

Snyder, B., Naseem, T., Eisenstein, J., & Barzilay, R. (2008). Unsupervised Multilingual Learning for POS Tagging. In *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1041–1050).

Sonai, K. A. M., Krishnan, J. K., Sayeed, M. S., & Muniapan, P. (2017). Comparison of Stochastic and Rule-Based POS Tagging on Malay Online Text. *American Journal of Applied Sciences*, *14*(9), 843–851. http://doi.org/10.3844/ajassp.2017.843.851

Soto, V., & Hirschberg, J. (2018). Joint Part-of-Speech and Language ID Tagging for Code-Switched Data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching* (pp. 1–10).