Khasi Shallow Parser

Medari Janai Tham
Computer Science & Engineering and Information Technology
Assam Don Bosco University
medaritham16@gmail.com

Abstract

A shallow parser is constructed for Khasi an Austro-Asiatic language, where noun phrases and verb phrases are identified. Formulation of what constitutes a Khasi noun phrase or verb phrase is carried out manually and marked in a corpus consisting of words already tagged with their corresponding part-of-speech tags. The parser is the first of its kind for the language, where the training corpus comprises of 24,194 chunks of noun and verb chunks out of a total of 3,997 sentences and 86,087 tokens. The approach in developing the parser is taken as a tagging problem using the Hidden Markov Model (HMM) and the results obtained have shown that a shallow parser is an appropriate first step when there is lack of information with regards to other phrases and the possible existence of lexical or syntactic mistakes in the training corpus.

1 Introduction

Shallow parsing is a process where a text is divided into non-recursive syntactical units such as noun phrases, verb phrases, etc. Non-recursive implies these phrases or chunks are non-overlapping and do not contain each other as proposed by Abney (1991). Shallow parsing has proven to be an alternative to full parsing of sentences, when only a subset of the information provided by a complete parser is sufficient for applications such as information retrieval, text summarization, etc. Further, it is also possible to enhance the corpus as and when information is available (Li and Roth, 2001), which is the current scenario with Khasi where only noun phrases and verb phrases have been annotated in the corpus. Khasi belongs to the Mon-Khmer branch of the Austro-Asiatic language family and spoken mainly in the state of Meghalaya. It is an analytic and partially agglutinative language having subject-verb-object (svo) word order, unlike the majority of Indian languages which are subject-object-verb (sov). It is not inflected and demonstrates derivational morphology in terms of affixes attached to a base word. These affixes can be easily detected and separated in any given word. The training corpus and test data are the same data set utilized in the development of a Khasi POS tagger (Tham, 2018) where the sentences have been supplemented with noun and verb phrases. The identified noun and verb chunks are non-recursive in nature and the actual constituents proposed for Khasi are given in section 3. Since the approach in constructing the parser is taken as a tagging problem, a Hidden Markov Model (HMM) part-of-speech (POS) tagger for Khasi with 95.68% accuracy (Tham, 2018) is also employed as a shallow parser, where only the data in the training corpus is altered to incorporate features relevant for parsing in the lines of Molina and Pla's (2002) shallow parser for English.

2 Related Work

Abney (1991) is credited with the introduction of the existence of chunks and he used hand crafted cascaded finite state transducers to detect chunks and clauses. However, Church's parser (1988) for detecting simple NPs can be attributed as the first statistical approach to noun phrase detection. Ramshaw and Marcus (1995) have approached chunking as a tagging problem by applying transformation based learning in detecting noun phrases and achieved recall and precision of 92% accuracy for base NP chunks while only 88% for tagging V and N chunks. Unlike Church's noun phrases (Church, 1988) which are simple, Ramshaw and Marcus

Proc. of ICON-2018, Patiala, India. December 2018 ©*2018 NLPAI*, pages 43–49

noun phrases include noun phrases formed with the use of conjunctions such as "and" and "or" or commas and where the possessive marker is treated as the first word of a noun phrase. Molina and Pla (2002) have also approached shallow parsing as another tagging problem by constructing what they termed as specialized HMMs where the learning and tagging procedures remain the same and adjustments have been made only to the training data and the output tags. These adjustments were carried out by varying the input and output combinations, and have reported results when there was an improvement from their baseline system comprising of the original training data and output tags. Their conclusion after comparing rule based systems, memory based systems, statistical systems, and combined systems was that combined systems gave better performances than individual systems with the exception of a Winnow system in some instances, but HMM systems also gave better performance than most systems, that too, requiring relatively less settings of information. Their overall performance reported an accuracy of $F_{\beta=1}$as 92.19.

In the Indian scenario, Singh et al. (2005) in their HMM chunker for Hindi, experimented with various tagging formats for chunk boundaries and chunk labeling and reported that for certain groups of words keeping only their POS tags improves accuracy than keeping both word and POS tag as tokens. They achieved a precision of 91.7% in chunk labelling. In the second contest conducted by International Joint Conference on Artificial Intelligence (IJCAI) on a Workshop on Shallow Parsing for South Asian Languages (Bharati and Mannem, 2007), POS and chunk annotated data comprising of 20,000 words of training data, 5000 words of development data and 5000 words of test data were provided for three Indian languages- Hindi, Bengali and Telegu. A total of eight teams participated using various approaches and PVS and Karthik (2007) is the only team that applies two learning techniques-HMM for chunk boundary detection and CRF for chunk labelling. They achieved best results in chunking for all the three languages Hindi, Bengali and Telegu with accuracy of 80.97%, 82.74% and 79.15% respectively. Tapping the morphological richness of a language, and justifying that a relatively small training corpus suffices, and avoiding the need of a large annotated corpus, a shallow parser for Marathi, Gune et al. (2010) achieved 97% accuracy for chunk identification using a 20,000 word size corpus. However, a recent noun phrase chunker for Marathi (Pawar et al., 2015) employed a CRF classifier for chunking. In this chunker, citing the lack of natural language processing (NLP) resources in India coupled with the fact that Marathi is a highly agglutinative language; the training corpus was generated automatically using Distant Supervision framework where the data is labeled according to some heuristic rules based on corpus statistics. Their reported F1 measure is 88.72%.

3 Labeling Khasi noun and verb chunks

According to Abney (1991), "a chunk consists of a single content word surrounded by a constellation of function words, matching a fixed template". As mentioned earlier, this imply chunks that are non-overlapping and do not contain each other. In this study, only noun and verb chunks have been identified. The elements of a Khasi noun chunk are similar to the noun phrases put forward by Jyrwa (1989) without the post-modifiers, while a verb chunk is taken to be the main verb itself along with any pre-modifiers such as auxiliary verbs excluding post-modifiers such as adverbs. The noun chunks excludes pronouns in the lines of Abney's (1991) definition of a chunk where pronouns are treated as orphans, and secondly, because they can also function as pronominal markers and subject enclitic (Jyrwa, 1989), for in such instances they do not syntactically function as noun chunks.

The corpus used for labeling the noun and verb chunks is a corpus annotated with part of speech tags from the BIS tagset for Khasi (Tham, 2018). The BIO labeling specified in Ramshaw and Marcus (1995) is followed for Khasi where each alphabet symbolizes the following:

B-XX: label **B** for a word starting a chunk of type XX.

I-XX: label **I** for a word inside a chunk of type XX.

O: label **O** for a word outside of any chunk.

Issues that surfaced while labeling both noun and verb chunks are highlighted below:

Basic noun phrase and the inclusion or exclusion of adjectives: According to Jyrwa (1989), the most basic noun phrase comprises of a number/gender marker also known as pronominal marker (PM), and a noun word followed by a subject enclitic (SE) as shown in example 1. Most of the abbreviations used in all the examples are in accordance with the Leipzig glossing rules[1], except when mentioned accordingly. In Khasi, pronominal markers are mandatory except in few instances where they are dropped. In example 1 the basic noun phrases present in the sentence are - *ka Banri ka*, and *ja*. However, during labeling a noun chunk, the subject enclitic has been excluded and the noun chunk is labeled up to and including the head noun, and we are left with *ka Banri* as a noun chunk. This is analogous to English noun chunks which can contain determiners and adjectives as specified in Sang et al. (2000). In Khasi, adjectives can occupy different positions in a sentence and they are included in a noun chunk only if they precede or immediately follow the noun they modify as shown in example 2 and 3. Therefore the possible pre-modifiers included in a Khasi noun chunk are demonstratives, cardinal numbers, quantifiers, pronominal markers, distributive particles, and adjectives (Nagaraja, 1985; Jyrwa 1989).

As mentioned earlier, instances where a pronominal is dropped are in vocative sentences, locative phrases and when a noun follows a verb (Jyrwa, 1989; Tham, 2018). In such cases a noun chunk comprises of the noun word without a pronominal marker.

1. ka Banri ka bam ja

   PM Banri SE eat rice

   “Banri is eating rice”

2. u diengsohphan bah

   PM jackfruit massive

   “a massive jack fruit”

3. long kaba skhem jingmut

   be REL strong mind

   “be strong minded”

Collocations of two or more nouns are part of the same noun chunk: Collocation of two or more nouns is a common phenomenon in Khasi where the actual meaning is derived from the summation of the words such as example 4. In most instances the noun(s) act as post modifier (example 5) while in some instances it acts as pre modifier(example 6). They are therefore labeled under the same noun chunk. It may be noted, that verbs tagged as nouns contribute to such collocations and hence give rise to noun chunks. Corpus analysis reflect that when a verb follows a noun it naturally becomes an element of the noun phrase comprising the noun in question as seen in examples 7 and 8, otherwise it is recommended that punctuation in the form of a comma (,) separates the noun and the following verb (example 9). However, when stylistic writing comes into play in the form of repetitions, then punctuation is not necessary as in example 10 where *sharai* a verb follows the noun *khynnah* and repeated as a stylistic element after the noun *blang*, but its attachment is to the noun *khynnah* and not the noun *blang*.

4. ka bai synniang kur

   PM cost fee clan

   “clan donation”

5. ka shuki dieng

   PM chair wood

   “wooden chair”

6. kynja kam

   type work

   “type of work”

7. shympriah thoh shun

   finger write lime

   “index finger”

8. sngi pdiang khatduh

   day accept last

   “last day of acceptance”

[1] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

9. na une u lum, iohi baroh

   from this PM hill, see all

   sawdong

   around

   "from this hill, we can see all around"

10. khynnah sharai blang sharai masi

    youth serve sheep serve cows

    "shepherd"

Labeling imitative noun chunks: Imitative words are group of words where the ancestor(s) and its successor(s) are associated phonetically in their pronunciation, and they are used more for their stylistic characteristic. When the ancestor and the successor are preceded with their own pronominal marker, then both are tagged as separate noun chunks (example 11), but instances where the pronominal marker occurs only before the ancestor and not the successor, then the phrase is taken as one noun chunk (example 12). Here the POS tags attached to each word are in accordance with the BIS tagset for Khasi (Tham, 2018).

11. ka shnong ka thaw

    "village"

    ka/PR_PRP_M/B-NP shnong/N_NN/I-NP ka/PR_PRP_M/B-NP thaw/N_NN/I-NP

12. ki per soh per syntiew

    "orchard"

    ki/PR_PRP_M/B-NP per/N_NN/I-NP soh/N_NN/I-NP per/N_NN/I-NP syntiew/N_NN/I-NP

Inclusion and exclusion of the conjunction bad in a noun chunk: The conjunction *bad* is comparable to the English conjunction "and" and can also participate as an element in a noun phrase. In example 13 the conjunction is part of the noun chunk, but in example 14 it is excluded from the noun chunk because the pronominal marker precedes the second noun, indicating that acceptable pre-modifiers of noun chunks are the ones mentioned earlier without overlapping.

13. i mei bad papa

    "mother and father"

    i/PR_PRP_M/B-NP mei/N_NN/I-NP bad/CC_CCD/I-NP papa/N_NN/I-NP

14. i mei bad i papa

    "mother and father"

    i/PR_PRP_M/B-NP mei/N_NN/I-NP bad/CC_CCD/O i/PR_PRP_M/B-NP papa/N_NN/I-NP

Possessive particle la labeled as an element of a noun chunk: One of the functions of *la* is as a possessive marker (Tham, 2018), and in the training corpus it has been labeled as a member of a noun chunk because syntactically when a noun phrase is the object of a preposition *la* can occur as the first element of a noun phrase (example 15) and the same can be said of *la* when the noun phrase is the object of a verb (example 16).

15. ban wad jingiada na la ki

    to seek protection from POSS PM

    briew

    person

    "to seek protection from his own people"

16. ka kyrngah la ka khlieh

    3SGF shook POSS PM head

    "she shook her head"

Basic verb phrases: The various forms of verb phrases present in the corpus are as follows.

- A basic verb phrase can comprise only of the main verb or can also include any preceding auxiliaries. For eg. *bam* (eat) or *la bam* (have eaten).

- Instances where only an auxiliary verb exists without the main verb, the verb chunk includes only the auxiliary verb. For e.g. *long* in example 17.

- Any two consecutive verbs are taken as two separate verb chunks. For e.g. *sdang hap* (starts falling).

- The infinitive phrase comprising of the infinitive *ban* (to) up to the main verb which may include auxiliaries in between is considered as a separate verb chunk as shown in example 18.

- The inclusion and exclusion of the conjunction *bad* as part of a verb chunk is in the lines of how noun chunks include the conjunction mentioned earlier.

The rest of the tokens present in the corpus that are outside the mentioned chunks are marked with the **O** tag.

17. ki long ki briew kiba bha

    3PL are PM person REL good

    "they are good people"

    ki/PR_PRP/O long/V_VAUX/B-VP
    ki/PR_PRP_PM/B-NP
    briew/N_NN/I-NP kiba/PR_PRL/O
    bha/JJ/O

18. ka la nang ban shad

    2SGF AUX knows to dance

    "she knows how to dance"

    ka/PR_PRP/O la/V_VAUX/B-VP
    nang/V_VM/I-VP
    ban/V_VAUX_VINF/B-VP
    shad/V_VM/I-VP

4 HMM shallow parser for Khasi

Following the work of Molina and Pla (2002), where shallow parsing is considered as a tagging problem, a standard HMM algorithm has been employed in developing an HMM Shallow Parser for Khasi. Molina and Pla (2002) have put forward a specialized HMM where alterations have been made in the training corpus while the training and tagging procedure remains intact. They have attained results at par with existing approaches especially when lexical information is added and achieved best $F_{\beta=1}$ as 92.23. Similarly the Khasi HMM POS tagger (Tham, 2018) has been used as a parser where changes have been made only in the training corpus.

Statistically, given a set of input symbols I and a set of output symbols C, tagging a sentence S=s_1,s_2... s_n of n symbols where $s_j \in I \; \forall \; s_j$with output tags c_1,c_2... c_n where $c_j \in C \; \forall \; c_j$ is given by

$$argmax_C \prod_{i=1}^{n} P(s_i|c_i)P(c_i|c_{i-1} \dots c_{i-k}) \qquad (1)$$

Taking into account Markov's assumptions a second order Markov model reduces equation 1 to equation 2.

$$argmax_C (\prod_{i=1}^{n} P(s_i|c_i)P(c_i|c_{i-1}, c_{i-2})) \qquad (2)$$

The probabilities are then estimated from the training corpus using maximum likelihood estimation, and linear interpolation has been carried out to counter any data sparsity problem encountered as shown in equation 3.

$$P(c_i|c_{i-2}, c_{i-1}) = \lambda_3 \hat{P}(c_i|c_{i-2}, c_{i-1}) + \lambda_2 \hat{P}(c_i|c_{i-1}) + \lambda_1 \hat{P}(c_i) \qquad (3)$$

Here $\hat{P}(c_i|c_{i-2}, c_{i-1})$, $\hat{P}(c_i|c_{i-1})$, $\hat{P}(c_i)$ are the trigram, bigram, unigram probabilities respectively, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

Further, Brants (2000) deleted interpolation is used for evaluating the λs and the Viterbi algorithm (Rabiner, 1989) is utilized to ensure an optimal path is taken when selecting the sequence with the highest probability.

In this analysis, since no more than noun and verb chunks have been identified and the BIO tagging scheme is utilized, the total number of chunk tags (output symbols) is 2n+1 i.e. 5 where n is the number of chunks. On the other hand, the input symbols involved words and their corresponding POS tags leading to a huge set of symbols. As suggested by Molina and Pla (2002) a specialization function f_s can be applied on the manually tagged training data T to produce a new training data $\check{T}$. Here the specialization function (equation 4) transforms every training pair <s_i,c_i> to <$\check{s}_i$, $\check{c}_i$> and therefore it changes the set of input symbols to $\check{I}$ and the output symbols to $\check{C}$.

$f_s : \mathrm{T} \subset (I \times C)^* \rightarrow \check{T} \subset (\check{I} \times \check{C})^*$ (4)

4.1 Testing and Evaluation

The transformations carried out for Khasi is accomplished by changing the input and output used for training. Initially for the baseline tagger only the POS tags are maintained as input symbols and the chunk tags as output symbols i.e. the training pair is $<p_i,c_i>$ where p_i is a POS tag and c_i the chunk tag. A sample input and sample output of the baseline HMM tagger is shown in example 19. The results of the tagger are then taken as the baseline results for the system (Table 1). In the next step, the POS tags are once more taken as input symbols, but the output symbols comprises of both POS tags and chunk tags (concatenated together with the period (.)). The new training pair is therefore $<p_i,p_i.c_i>$. Example 20 shows a sample output of the HMM shallow parser on the same input shown in example 19. The test data consist of 402 sentences which includes 2,210 noun and verb chunks and the tagging results are shown in Table 1, indicating that adding just POS information to the chunk category has dramatically improved the accuracy to $F_{\beta=1}$as 95.51 as compared to the baseline of $F_{\beta=1}$as 86.94.

19. V_VM RB IN N_NN N_NN RD_PUNC (input)

    V_VM/B-VP RB/O IN/O N_NN/B-NP N_NN/I-NP RD_PUNC/O (output)

20. V_VM/V_VM.B-VP RB/RB.O IN/IN.O N_NN/N_NN.B-NP N_NN/N_NN.I-NP RD_PUNC/RD_PUNC.O (output)

The individual results for noun and verb chunks are given in Table 2 and analysing the results reveals that in most cases where the chunks were not detected accurately are mainly due to the following:

- When the noun chunk is the object of the preposition and the chunk contains an adverb as the first element then it fails to identify the adverb as the starting element of the noun chunk.
- Non detection of conjunctions which are part of a noun chunk.
- As mentioned in section 3 consecutive nouns are always considered as part of the same noun chunk, but in some instances this is not true. These phrases are semantically determined, which is difficult to detect at this stage of the parser. For instance in example 21, *shipara* and *kynthei* are not part of the same noun chunk, but the tagger has placed both of them within the same noun chunk.

21. ar ngut ki khynnah shipara kynthei

    two CLF PM youth sibling girl

    bad shynrang

    and boy

    "two siblings, a girl and a boy"

- Auxiliary verbs following another auxiliary verb tend to be tagged as part of a new verb chunk when they are actually elements of the previous verb chunk.

	Precision	**Recall**	$F_{\beta=1}$
Baseline	86.38%	87.51%	86.94
Khasi Shallow Parser	94.39%	96.65%	95.51

Table 1: Results

Chunk	**Precision**	**Recall**	$F_{\beta=1}$
NP	93.4%	97.23%	95.28
VP	95.34%	96.1%	95.72

Table 2: Noun and Verb chunk results

5 Conclusion

This work has initiated a corpus of noun and verb chunks, and an HMM shallow parser for Khasi which requires Khasi text tagged with their part of speech. The details of what constitutes a noun chunk or a verb chunk were highlighted keeping in mind that identifying a Khasi noun chunk or a verb chunk from a given text is a new initiative for the language. The results of the parser are encouraging and are in

parity with what is reported for English and other Indian languages and enhancing corpus size will facilitate further testing. In future, when analysis of other Khasi phrases is available, what will remain is incorporating the acquired information only in the corpus without the need of modifying the tagging algorithm.

References

Steven Abney. Parsing by Chunks. 1991. In R. Berwick, S. Abney and C. Tenny (Eds), *Principle–based Parsing*. Kluwer Academic Publishers. MA (pp 257-278).

Thorstens Brants. 2000. TnT-A statistical part of speech tagger. *In Proceedings of the Sixth Conference on Applied Natural Language Processing* (pp. 224-231). Seattle, Washington. doi:10.3115/974147.974178.

Akshar Bharati and Prashanth R. Mannem. 2007. Introduction to Shallow Parsing Contest on South Asian Languages. *In Proceedings of the IJCAI and the Workshop On Shallow Parsing for South Asian Languages (SPSAL)* (pp. 1-8). Hyderabad. Retrieved from https://pdfs.semanticscholar.org/448a/80b1d27ad563d494fd698a77dfe09bd67bdf.pdf

Kenneth Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. *In proceedings of Second Conference on Applied Natural Language Processing* (pp. 136-143). Austin, Texas. doi: 10.3115/974235.974260

Harshada Gune, Mugdha Bapat, Mitesh M. Khapra, and Pushpak Bhattacharyya. 2010. Verbs are where all the action lies: experiences of shallow parsing of a morphologically rich language. *In Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 347-355). Beijing. Retrieved from http://www.aclweb.org/anthology/C10-2040

Mumtaz Bory Jyrwa. 1989. *A Descriptive study of the Noun Phrase in Khasi* (Doctoral dissertation). Retrieved from http://shodhganga.inflibnet.ac.in/handle/10603/61398

Xin Li and Dan Roth. 2001. Exploring evidence for Shallow Parsing. 2001. *Proceedings of the Conference on Computational Natural Language Learning*. Toulouse. doi:10.3115/1117822.1117826

K.S. Nagaraja. 1985. *Khasi a Descriptive Analysis* Pune, India. Deccan College Post-Graduate & Research Institute.

Antonio Molina and Ferran Pla. 2002. Shallow Parsing using Specialized HMMs. *Journal of Machine Learning Research. Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing*. 2, 595-613. Retrieved from http://jmlr.org/

Sachin Pawar, Nitin Ramrakhiyani, Girish K. Palshikar, Pushpak Bhattacharyya, and Swapnil Hingmire. 2015. Noun Phrase Chunking for Marathi using Distant Supervision. *In Proceedings of the 12th International Conference on Natural Language Processing* (pp. 29-38). Trivandrum. Retrieved from http://www.aclweb.org/anthology/W15-5905

Avinesh PVS and G Karthik. 2007. Part-of-speech tagging and chunking using conditional random fields and transformation based learning. *In Proceedings of the IJCAI and the Workshop On Shallow Parsing for South Asian Languages (SPSAL)* (pp. 21-24). Hyderabad. Retrieved from https://pdfs.semanticscholar.org/448a/80b1d27ad563d494fd698a77dfe09bd67bdf.pdf

Lawrence R. Rabiner. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *In Proceedings of the IEEE*. *77*(2), 257-285. doi: 10.1109/5.18626.

Tjong Kim Sang, Erik F. and Sabine Buchholz. Introduction to the CoNLL-2000 Shared Task: Chunking. 2000. *In Proceedings of CoNLL-2000 and LLL-2000* (pp. 127-132). Lisbon. doi:10.3115/1117601.1117631

Akshay Singh, S M Bendre and Rajeev Sangal. 2005. HMM chunker for Hindi. *Proceedings of IJCNLP*: *The Second International Joint Conference on Natural Language Processing*, Jeju Island. Retrieved from http://aclweb.org/anthology/I05-2022

Medari J. Tham. 2018. Challenges and Issues in Developing an Annotated Corpus and HMM POS Tagger for Khasi. *The 15th International Conference on Natural Language Processing.* (forthcoming). Patiala, Punjab.