Improved Neural Word Segmentation for Standard Tibetan

Collin Brown
Indiana University
coljbrow@iu.edu

Abstract

As Tibetan is traditionally not written with word delimiters, various means of word segmentation are necessary to prepare data for downstream tasks. Neural word segmentation has proven a successful means of parsing Tibetan text, but current performance lags behind that of neural word segmenters in other languages, such as Chinese or Japanese, and even behind languages with relatively similar orthographic structures, such as Vietnamese or Thai. We apply methods that have proven useful for these latter two languages toward the development of a neural word segmenter with the goal of raising the peak performance of Tibetan neural word segmentation to a level comparable to that reached for orthographically similar languages.

Keywords: Tibetan, Word Segmentation

1. Introduction

Tibetan is a language—or rather, a number of languages and dialects of varying degrees of mutual-intelligibility—spoken in Tibet, a region overlapping a number of provinces in modern-day China including the Tibetan Autonomous Region, Sichuan, and Qinghai. Diaspora communities reside also in India, Nepal, and Bhutan; and a substantial, if smaller, number live also in Switzerland, Canada, the United Kingdom, and the United States (among many other countries).

Tibetan belongs to the Sino-Tibetan language family and is traditionally placed in the Tibeto-Burman branch, though the phylogeny of the family remains hotly contested. The Tibetan family can be further divided into various dialect and language groups, including Central (or Ü-Tsang) with approximately 1.2 million speakers, Amdo with 2.5 million, and Khams with 2 million, among others (Eberhard et al., 2024). However, Standard Tibetan generally serves as a lingua franca among them; thus, expanding the resources available to the language provides benefits not only to native speakers but also to the broader Tibetan community, whatever their regional or dialectal background. By improving word segmentation for Tibetan, we hope to facilitate the creation of further tools—word prediction, sentiment analysis, etc—which might make the language easier for its speakers to use in the digital domain, easing linguistic pressures that motivate them to switch to languages with more support, such as English, Mandarin Chinese, or Hindi.

Many Asian scripts are not written with spaces between words, and this obviously presents certain problems when one wishes to engage in most computational tasks, the models for which tend to operate on words rather than characters. Standard Chinese, another such space-less language, has the benefit of using characters, each of which are semantically heavy; character-embeddings allow for the training of highly accurate models. While some languages, such as Korean, have broadly adopted the practice of placing spaces between words, many orthographies descended from either the Indic or Sinitic traditions continue to go without them. Furthermore, some languages make use of orthographic features that make word segmentation an easier task; Japanese in particular uses multiple different scripts, and the transitions between these often serve as strong indicators of word boundaries. However, we do not have such luxuries with Tibetan which only explicitly marks syllable and sentence boundaries. While researchers have reached an accuracy of upwards of 98% for Japanese texts (Kitagawa and Komachi, 2018) and 97% for Chinese (Cai et al., 2017), Tibetan lags behind. Duanzhu et al (2020)

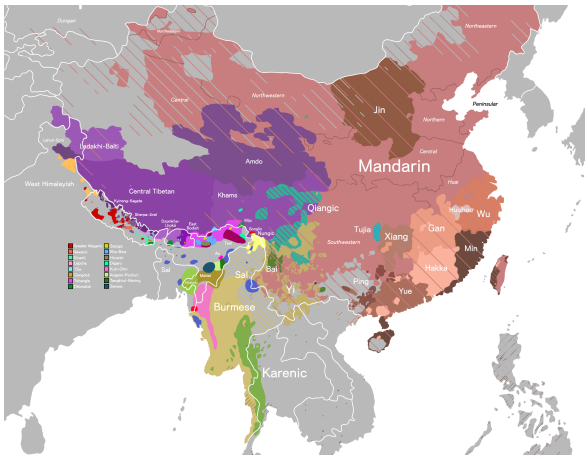


Figure 1: The Tibetic family appears here in shades of purple in the northwest quadrant. GalaxMaps, CC BY-SA 4.0, via Wikimedia Commons.

2nd *EURALI Workshop*, pages 12–17
25 May, 2024.

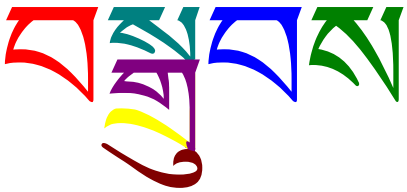
report a binary accuracy of 93.4% with an f1-score of 94.11% and a recall of 94.2%; Wang & Yang (2018), a f1-score of 94.1% and a recall of 93.89%; and Li et al (2022) an f1-score of 92.31% (Duanzhu et al., 2021; Wang and Yang, 2018; Li et al., 2022).

While the success of Chinese and Japanese can to some degree be attributed to the vast resources available for these two languages, neural word segmentation research for smaller (though by no means, small) languages such as Vietnamese and Thai have reached an accuracy of around 96% or higher (Zheng and Zheng, 2022).

2. Background

Phonetically, the Tibetan syllable is of only moderate complexity, but the language's standard orthography preserves the highly complex syllable structures of the ninth and tenth centuries. While spoken syllables in Lhasa Tibetan may begin and end with at most a single consonant respectively, they may be written with upwards of four initial consonants and two final consonants. Furthermore, vowels are not written as distinct letters but instead added as diacritics above the "head" letter, or the letter whose phonetic value serves as the basis for the onset of the syllable.

The maximally complex Tibetan syllable is composed of a prescript letter, a head letter, a postscript and a post-postscript letter. All but the head letter are (usually) composed of a single, simple letter, but the head letter can itself be composed of a superscript, a root, upwards of two subscripts, and a vowel diacritic.



ད (b), ས (s), ག (g), ར (r), ུ (u), ད (b), and ས (s)

Figure 2: A maximally complex Tibetan syllable. The past tense form of the word, གྲུབ (sgrub), meaning "to complete".

As can be seen in figure 2, the unique complexity of Tibetan syllables allows them to carry a relatively high degree of semantic value; thus, they can serve as stronger indicators of word boundaries than syllables in more orthographically shallow languages.

There are thirty standard letters that may serve as the root of the head of a syllable. Onto these, four subscripts may be attached—ྱ (y), ྲ (r), ྭ (l), ྭ (w)—but only certain head-subscript combinations are allowed. In addition, ྲ (r) and ྭ (w) may appear together on the same root, meaning that the total number of root-subscript combinations comes out to fifty-five unique arrangements. Onto these, one may attach three superscripts—ར (r), ལ (l), ས (s)—and again these are only allowed in particular arrangements, meaning that the total number of head letters which feature a unique superscript, root, and subscript is only thirteen. Adding these, as well as the unique superscript-root combinations, to our running total gives us one-hundred-and-one unique head letters. Each of these may take up to one vowel diacritic, of which there are four, yielding six-hundred-and-six unique head letters. These diacritics are ི (i), ུ (u), ེ (e), and ོ (o).

Onto these, one may add some combination of prescript and postscript letters. There are ten postscript letters and two post-postscript letters (though really only one, as the other has been dropped in most writing). The post-postscript letter may only appear after four of the postscript letters, meaning there are a total of fourteen possible postscript combinations. Confusingly, one of these postscript letters, འ may also carry a vowel diacritic, though it is usually limited to འི (i) or འུ (u); however, འེ (e) and འོ (o) do appear, albeit rarely. Thus, we have eighteen possible postscript combinations.

Despite there being a maximum of one prescript letter, calculating the number of possible combinations is less straightforward given that there are more restrictions on which letters may appear in certain positions. There are five prefixes—ག (g), ད (d), བ (b), མ (m), and འ (a)—and calculating the total number of unique prescript-head letter combinations created by them is quite difficult given their distribution. Disregarding super- and sub-scripts, as well as vowels, there are a total of fifty-three unique prescript-head letter combinations. If we are liberal with our estimates, we would say that the number of unique prescript-head letter combinations (including all our superscript, subscript, and vowel combinations) comes out to around three thousand unique combinations. Assuming that many of these do not appear in the actual written language, we might lower this down to only a couple thousand unique combinations, onto which we would then necessarily add our various postscript letters, bringing out estimated total number of unique syllables into the tens of thousands.

This number is highly misleading, as we find out when we compile a dictionary of all the sylla-

bles that appear in any particular Tibetan corpus. The true number of unique syllables to be found in actual texts is considerably lower, usually in the sub-ten-thousand range, and if we filter out those that appear less than five times—as we do in our model—we arrive at much more modest numbers, usually between two and five thousand unique syllables, depending on the size and variety of the corpus.

In any case, the semantic load of the Tibetan syllable, as well as the fact that the vast majority of word boundaries are also syllable boundaries, allows us to use syllable embeddings as a heuristic by which to train our model.

In standard, written text, all syllables are delimited by a unique punctuation mark known as the *tseg*, written '་'. This allows us to easily parse through a text and separate out each syllable, whereas many other languages that make use of syllable embeddings for neural word segmentation—such as Thai or Khmer—must engage in more complex syllable-identifying methods beforehand. While certain questions do arise about what constitutes a word-boundary in Tibetan, for our purposes we may treat word boundaries as a subset of syllable boundaries. Specifically, the genitive and agentive cases sometimes take the form of postscripts on the final syllable of words ending in vowels; in the spoken language, they are realized via umlaut or lengthening of this final vowel, and so we will treat them as part of the word rather than separate particles.

3. Corpus

While the most extensive corpora available for Tibetan are limited to Classical Tibetan, we were able to make use of the UVA Tibetan Spoken Corpus (Germano et al., 2017) which, while a couple decades old, represents the most easily accessible corpus of pre-segmented text available. This corpus was compiled by the *Tibetan and Himalayan Digital Library* project which is affiliated with the University of Virginia and reflects the colloquial language of people living in Tibet rather than the highly formal, literary language often found in religious and official texts. As we intend to apply this word segmentation model towards the development of tools aimed at making Tibetan more accessible in the digital realm, it was important that the corpus reflect the kinds of language used by everyday people.

The corpus indicates word boundaries with a space, meaning that with minimal processing we can clean the corpus of non-Tibetan text, and divide it into syllables, storing each with some indication as to whether or not it is word-final. With 807,033 total syllables, we can take those which appear with a frequency of at least 5, resulting in 2584 embeddings. In total, about 18.52% of the syllables in this corpus are non-word-final (meaning they don't mark a word boundary). This is quite a bit smaller than the corpus used to train the AttaCut model, from Chormai et al (Chormai et al., 2019), which featured 2.56 million syllables. Similarly, whereas Duanzhu et al (2020) employ a corpus with 160,000 sentences, ours features only 70,000 (Duanzhu et al., 2021).

Unsurprisingly, the most common syllables found in the corpus include ཡ, an incredibly common nominalizer and derivational suffix; ལ, the oblique case particle; དང, a conjunction and comitative / associative particle; ཀྱི, the medial demonstrative; and the various case and TAM endings and particles that compose Tibetan's robust nominal and verbal systems.

4. Methods

A somewhat recent and effective method for neural segmentation of Thai words is the use of syllable embeddings as input features. Training a neural network to identify word boundaries based on syllable embeddings rather than characters has proven quite effective, as evidenced by the AttaCut model developed by Chormai et al (Chormai et al., 2019). Because Tibetan explicitly marks syllable boundaries, and because of the orthographic depth of the language—with a spelling system that preserves pronunciations from antiquity—we determined that it might be particularly useful in improving performance.

In a manner similar to the AttaCut model, Nguyen 2019 makes use of BiLSTM to generate embeddings for the syllables present in a Vietnamese corpus and uses these to train a model to determine word-boundaries with a 98% accuracy. However, this model's success comes in part due to its use of a rule-based word segmenter, RDRsegmenter, as outlined in (Nguyen et al., 2018), in conjunction with its neural method. RDRsegmenter produces a set of word-boundary tags whose embeddings are concatenated with each syllable's embedding to produce those that are used to train the final model (Nguyen, 2019; Nguyen et al., 2018). Similar methods have proven effective for Chinese since each character corresponds, with some exceptions, to one syllable and one morpheme (Qian and Liu, 2012). It should be the case that Tibetan, which adheres less strictly to this one-syllable-one-morpheme structure, can still benefit from the application of this method.

Some combination of syllable embeddings, character embeddings, and word-boundary embeddings generated by a rule-based model have

proven useful for word segmentation in many of the languages of East and Southeast Asia, including Khmer, Chinese, and even Classical Tibetan, for which there exists more readily available corpora owing to the digitization of many Buddhists texts (Buoy et al., 2020). Given the intense conservatism of Tibetan orthography it may be possible to supplement a corpus of modern, standard Tibetan with texts from Classical Tibetan; however, initial tests yielded no benefits. More research is required to determine if this is a viable route for improvement.

A last note worth considering is the presence of non-standard text within the corpus. Certain sequences, especially in older texts, lack the syllable delimiter, complicating the pre-processing necessary for our model. In future research, it may be worth considering the implementation of a syllable segmenter which would insert syllable boundaries where they are stylistically omitted from text (or left out in error). Furthermore, it would be necessary to operate on a sub-syllable basis if one wished to separate certain instances of various cases which modify words at such a level. For example, when a noun's final syllable lacks any post-script letter, the genitive case takes the form, -འི (-i), which is given no special treatment here but which may, in other applications, necessitate further delimiting.

5. Implementation

By generating syllable embeddings via a Word2Vec model, we are able to train a model to predict the probability that a given syllable—the center of our context window—is non-word-final. Word2Vec is used in order to ensure manageable model size. We limited our syllable embeddings to only those syllables which appeared at least five times in the corpus; this may present an issue when our model is faced with a much larger corpus with many more unique syllables, some of which may appear semi-frequently, as well as when presented with regional or alternative spellings. To help with this and to push our performance past 96% accuracy, it may prove useful to implement a rule-based segmenter as done in Nguyen 2019, whose predictions should improve the accuracy of our neural segmenter when coupled—or rather concatenated—with our syllable embeddings. Furthermore, we hope to find a larger corpus on which to train our model in order to reduce the number of out-of-vocabulary syllables our model must cope with.

Early attempts at Tibetan word segmentation drew on MaxMatch algorithms and rudimentary statistical models, but with the proliferation of neural networks throughout natural language processing, the task has largely adopted such methods.

Hyperparameters	
Embedding Size	400
Learning Rate	1e-5
No. Layers	5
Window Size	3
Batch Size	64
Epochs	20
Accuracy:	96.87%

Table 1: The above hyperparameters are shown to approach those optimal for the model.

Drawing on a similar method to Liu et al (2015), another implementation of neural Tibetan word segmentation, we implement our model as a binary decision task, with the model labelling each syllable as either word-final or non-word-final (Liu et al., 2015). Syllable delimiters and any word delimiters are removed. Unlike Duanzhu et al (2021), we treat each syllable as an irreducible unit; they implement character embeddings in addition to syllable embeddings, which proves useful for certain purposes such as morphological analysis but introduces more opportunities for error (Duanzhu et al., 2021). We opt for a more straightforward model, considering only discrete syllables within a context window, and maintain a simple binary output rather than the more complex tag sets used in some implementations, such as Liu et al (2015) and Wang & Yang (2018) (Liu et al., 2015; Wang and Yang, 2018).

Currently, the vast majority of corpora are available not in Modern Tibetan but instead in Classical Tibetan, due to the many Buddhist texts that have been digitized from that period. Compiling a larger corpus in Modern Tibetan would provide our model with more data and reduce the instances of out-of-dictionary syllables. Initial tests involving the training of a model on The Annotated Corpus of Classical Tibetan (ACTib) (Meleen and Roux, 2020), followed by fine-tuning on a Standard Tibetan corpus, proved unsuccessful in yielding benefits compared to training solely on Standard Tibetan.

As we can see from figure 3, binary accuracy is not improved by expansion of the window size beyond one syllable on either side of the target. We might have assumed that a broader window would allow the model to differentiate between the occurrence of certain common syllables in various contexts, especially in words with more than two or three syllables, but this does not seem to be the case. Rather, as figure 4 reveals, performance is much more contingent on embedding size. This is somewhat unsurprising; in a language such as Tibetan where such units often correspond with morphemes, much meaning may be packed in.

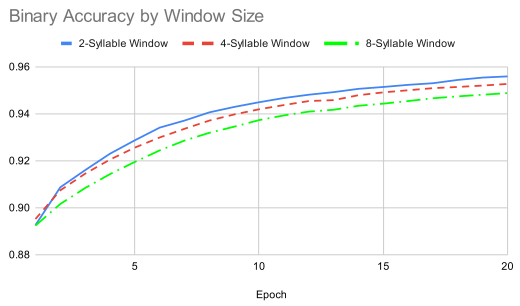


Figure 3: The window size of the model yields the best performance when only accounting for one syllable on either side of the target. Greater widths yield worse results overall, indicating that a local, relatively simplistic system can account for most word boundaries.

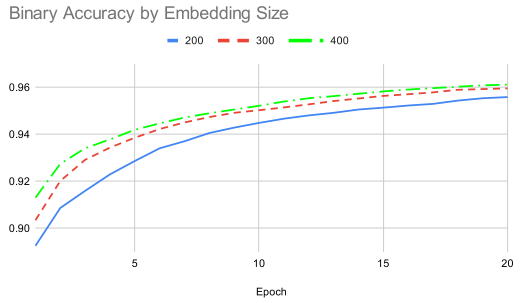


Figure 4: Larger vector sizes yields greater performance, but this diminishes above a value of 300. With model size in mind, we determine that values above 400 do not yield sufficient returns.

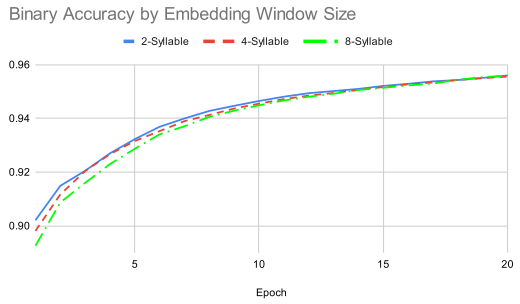


Figure 5: By changing the window we use when training our embeddings, we find a similar, albeit less pronounced, effect as with the model's window. Here, a 3-syllable window is considered optimal.

6. Results

While unable to achieve a level of performance found in more resource-rich languages, we are able to match that of languages with similar orthographic traditions. Furthermore, our findings indicate a great margin for improvement via the application of rule-based heuristics and larger corpus sizes.

If it were the case that the model primarily considered the target syllable, we would expect a 1-syllable window to outperform the 3-syllable window; however, this 3-syllable window outperforms both the 1- and 5-syllable window (and any greater number), indicating that it is the immediate, local context (and not any more-distant relation) that can account for most word boundaries. Augmenting this local window with more information (such as a rule-based heuristic) may yield further benefits.

The relatively great impact of embedding size does reflect the semantic weight of Tibetan syllables; their complexity provides information about word boundaries not contained in the orthographic units of even structurally similar languages. This is undoubtedly influenced as well by the syllable-to-morpheme ratio of Tibetan which (like many neighboring languages) tends to approach 1.0.

7. Conclusion

By implementing methods that have proven successful for neural word segmentation in orthographically similar languages, such as Thai and Vietnamese, we have been able to achieve a level of performance approaching the most performant word segmenters for Standard Tibetan, though further exploration may yield enough improvements so as to surpass the current peak performance and bring Tibetan word segmentation up to a comparable level as has been achieved for these other languages. Currently, we are limited by the availability of large corpora in Standard Tibetan; the acquisition of more data in addition to the refinement of existing methods and their augmentation with novel heuristics would, in benefitting neural word segmentation, provide downstream benefits for all varieties of natural language processing tasks.

8. Bibliographical References

Rina Buoy, Nguonly Taing, and Sokchea Kor. 2020. Khmer word segmentation using bilstm networks.

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. 2017. Fast and accurate neural word segmentation for Chinese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 608–615,

Vancouver, Canada. Association for Computational Linguistics.

Pattarawat Chormai, Ponrawee Prasertsom, and Attapol Rutherford. 2019. Attacut: A fast and accurate neural thai word segmenter.

Sangjie Duanzhu, Cizhen Jiacuo, and Cairang Jia. 2021. Revisiting tibetan word segmentation with neural networks. In *Chinese Lexical Semantics*, pages 515–524, Cham. Springer International Publishing.

David Eberhard, Gary Simons, and Charles Fennig. 2024. Ethnologue: Languages of the world.

David Germano, Edward Garrett, and Stephen Weinberger. 2017. Uva tibetan spoken corpus.

Yoshiaki Kitagawa and Mamoru Komachi. 2018. Long short-term memory for Japanese word segmentation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Yan Li, Xiaomin Li, Yiru Wang, Hui Lv, Fenfang Li, and La Duo. 2022. Character-based joint word segmentation and part-of-speech tagging for tibetan based on deep learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Huidan Liu, Congjun Long, Minghua Nuo, and Jian Wu. 2015. Tibetan word segmentation as subsyllable tagging with syllable's part-of-speech property. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 189–201, Cham. Springer International Publishing.

Huidan Liu, Minghua Nuo, Longlong Ma, Jian Wu, and Yeping He. 2011. Tibetan word segmentation as syllable tagging using conditional random field. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 168–177, Singapore. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Marieke Meleen and Élie Roux. 2020. The annotated corpus of classical tibetan (actib).

Dat Quoc Nguyen. 2019. A neural joint model for Vietnamese word segmentation, POS tagging and dependency parsing. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 28–34, Sydney, Australia. Australasian Language Technology Association.

Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. 2018. A fast and accurate Vietnamese word segmenter. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xian Qian and Yang Liu. 2012. Joint chinese word segmentation, pos tagging and parsing.

Lili Wang and Hongwu Yang. 2018. Tibetan word segmentation method based on bilstm$_{crfmodel}$. In *2018 International Conference on Asian Language Processing (IALP)*, pages 297–302.

Kexiao Zheng and Wenkui Zheng. 2022. Deep neural networks algorithm for vietnamese word segmentation. *Scientific Programming*, (8187680).