Modelling Tibetan Verbal Morphology

Qianji Di Ekaterina Vylomova Timothy Baldwin

The University of Melbourne, Melbourne, Australia
Parkville, Victoria 3010, Australia
qdi@student.unimelb.edu.au {ekaterina.vylomova,tbaldwin}@unimelb.edu.au

Abstract

The Tibetan language, despite being spoken by 8 million people, is a low-resource language in NLP terms, and research to develop NLP tools and resources for the language has only just begun. In this paper, we focus on Tibetan verbal morphology — which is known to be quite irregular — and introduce a novel dataset for Tibetan verbal paradigms, comprising 1,433 lemmas with corresponding inflected forms. This enables the largest-scale NLP investigation to date on Tibetan morphological reinflection, wherein we compare the performance of several state-of-the-art models for morphological reinflection, and conduct an extensive error analysis. We show that 84% of errors are due to the irregularity of the Tibetan language.

1 Introduction

Tibetan is the language used by Tibetan people, in Tibetan areas in China — Tibet Autonomous Region, Qinghai Province, Sichuan Province, Gansu Province, and Yunnan Province — as well as parts of Nepal, Bhutan, India, and Pakistan. Among the languages of the Sino-Tibetan language family and Chinese national writing systems, the Tibetan language has one of the longest histories and most extensive bodies of literature. Although Tibetan is roughly divided into three dialects (Weizang, Khampa, and Amdo), the written language is uniform across all regions (Jitaiga, 2018).

The earliest work on Tibetan information processing began in the 1980s, focusing on fonts, character encoding support, and text input methods. This early work resulted in the ability to develop and share digital text resources in Tibetan, which benefited Tibetan scholars and the Tibetan diaspora (Zhijie, 2009). Tibetan information processing technology is rapidly developing, but some key problems remain, including the analysis of Tibetan verbs. Recent advances in Tibetan NLP have included the ability to automatically identify verbs, analyze the rules governing word form changing processes, and reveal various linguistic phenomena of Tibetan verbs (Zhijie and Rangzhuoma, 2010; Zhijie, 2005).

In this paper, we specifically focus on the Tibetan verbal inflection system, which is known for its irregularity (Suonanjiancuo, 2013). Specifically, we make use of the morphological reinflection task recently introduced under the umbrella of SIGMORPHON (Cotterell et al., 2018). We train several state-of-the-art machine learning models for reinflection and provide an extensive error analysis. We find that the original experimental data contained some errors, and the models do not cater to idiosyncrasies of the Tibetan language. After correcting errors in the data, experimental results improve. We also develop a new dataset for Tibetan verbs comprising 1,433 verbal lemmas with their present, past, future, and imperative forms.[1]

2 Tibetan Language

The Tibetan language belongs to the Tibetan branch of the Tibeto-Burman language group of the Sino-Tibetan language family. The Tibetan script is an abugida or alphasyllabary, whereby consonant–vowel sequences are written as a single unit. There are two schools of thought regarding the origin of Tibetan literature: some scholars be-

[1] The dataset is available at https://github.com/victoriadqj/Tibetan-Verb-Lexicon.

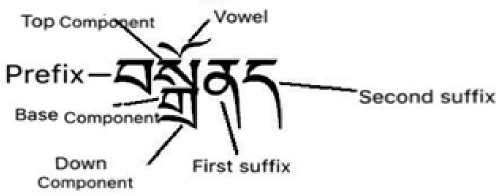

Figure 1: Lexicographic breakdown of a Tibetan word.

lieve that in the 7th century CE, the king Srongtsen Gampo in the Tubo era sent the Tibetan linguist Thombus Sangbu to North India to study Sanskrit, and that he created the Tibetan script based on Sanskrit (Wang, 1980). However, believers of the "Bonismo" religion hold that the Tibetan language evolved from Xiangxiong (Li et al., 2009).

Tibetan grammar is relatively rich, and verbs are inflected for tense and mood as follows: present, past, future, and imperative. Taking གཏབ *dep* "cut" as an example, the future form is གཏབ *dep* "cut", past form is གཏབ *dep* "cut", present form is གཏབས *di* "cut", and the imperative form is གཏབས *di* "cut"

In the Tibetan writing system, individual units (in the form of consonant–vowel sequences) are often referred to as "components". One or more components constitute one "character", which is monosyllabic. One or more characters form a "word". Each syllable in Tibetan has a base component, which is a consonant and determines the base pronunciation of the syllable. Vowel symbols can be added above or below the base component to indicate different vowel sounds, in the form of top components above the base component, and one bottom component below it. Sometimes there is a prefix, indicating that the syllable is consonant-initial. There can also be one or two suffixes after the base component, indicating that the syllable has one or two consonants in addition to the base consonant. Figure 1 provides an example of the composition of a Tibetan word.

3 Related Work

Verbs are the core and fundamental elements of Tibetan grammar.[2] In the "Tibetan Grammar Tutorial" (Jumian, 1982), the authors elaborate on written verbs in Tibetan, and propose about 1,300 written syllables. Around 70% of verbs in their data undergo tense inflection, and the other 30% are invariant under inflection; for only 20% of verbs is the imperative form different from the base lemma (Qu, 1985). Qu and Jing (2011) in "Sound Theory" defined categories of Tibetan letters, and elaborated on the composition of Tibetan verbs, principles of transitiveness, as well as tenses. Later, Jitaijia (2013) systematically studied functions of verbs in sentences and the relationship between verbs and other components in the sentence. Hill (2010) presented an overview of Tibetan verbal morphology. The author proposed to categorize Tibetan verbs into 11 paradigms, although there were still many irregular among frequently used ones. Still, scholars over the years have formed very different opinions on the morphosyntax of Tibetan.

Although we only consider Tibetan here, this work continues and further extends the work of Gorman et al. (2019). There, the authors conducted a detailed analysis of errors typically made by state-of-the-art morphological reinflection systems, in addition to introducing a novel error taxonomy that we utilize in this research.

4 Materials and Methods

4.1 The Task and Data

Following Gorman et al. (2019), we experiment with the morphological reinflection task (sub-task 1). The training data consists of triples of lemma, morphosyntactic features, and inflected form. In the test phase, the inflected form for an unseen lemma–morphological feature pair must be predicted.

We used the original datasets provided by Cotterell et al. (2018), which include two training sets for Tibetan, namely low-resource (100 samples) and medium-resource (158 samples).[3] Both test and development data comprise 50 samples. As part of this research, we develop a high-resource set of 1,433 instances.[4]

[2] This can be clearly concluded from the ancient Tibetan masterworks "Thirty Laud" and "Sound Theory" (Qu and Jing, 2011)

[3] All encoded in UTF-8, based on the standard Tibetan Unicode mapping which was released in 1991.

[4] The format follows the UniMorph annotation scheme (Sylak-Glassman et al., 2015).

Model	Low	Med	High
Lemma Copy	0.44	0.44	0.44
SMP Baseline	**0.54**	0.48	**0.50**
A&G 2017	0.34	0.46	0.48
M&C 2018	0.42	**0.52**	0.46

Table 1: Results over the original datasets (best in bold).

Model	Nonce	Allomorphy
Lemma Copy	12	17
SMP Baseline	8	18
A&G 2017	10	16
M&C 2018	16	15

Table 2: Absolute number of errors on the test set made by each system trained in medium-resource setting.

4.2 Systems

For our experiments, we consider four models: (1) a naive baseline, whereby we simply return the lemma as the inflected form ("Lemma Copy"); (2) the baseline model used in SIGMORPHON 2017/2018 shared tasks (Cotterell et al. (2017, 2018): "SMP Baseline"); (3) Aharoni and Goldberg (2017)'s hard attention neural model ("A&G 2017"); and (4) Makarov and Clematide (2018)'s neural transduction models ("M&C 2018"). The latter two models achieved the highest scores in low- and medium-resource settings in the SIGMORPHON 2017/2018 shared tasks. Both are essentially neural seq-to-seq models (developed using Dynet framework (Neubig et al., 2017)) that rely on the assumption of nearly-monotonic alignment between a lemma and its inflected form, and learn a sequence of edit operations to perform string transduction.[5] The SMP Baseline model is non-neural, and first aligns strings using Levenshtein distance, and then extracts prefix- and suffix-based transformations.

5 Results

The results obtained by the four models are shown in Table 1. After manually reviewing errors across the four systems, we found not only system errors, but also errors in the data. In error analysis we employed the error taxonomy proposed by Gorman et al. (2019) and identified the following types: (1) target errors in the dataset; and (2) prediction errors. We further break down prediction errors into: (2.1) nonce-word errors (where a model generates a word which clearly violates lexicographic or morphophonetic constraints in the language); and (2.2) allomorphy errors (where the wrong inflectional pattern is applied, i.e. a plausible inflection is generated, but it does not correspond to the indicated class).

5.1 Target Errors

Target errors are mainly due to errors in the Wiktionary source data[6] and incorrect extraction of paradigm tables, e.g. the lemma not existing in the lexicon, the inflected form not matching the indicated lemma, the positions of the inflected form and lemma being reversed, or even unrelated words appearing within a paradigm. Consider an example taken from the training data for the medium set. The lemma is སྒྲིག *zhegh* "arrange", the imperative form of which is said to be སྒྲིག *zhegh* "arrange". In practice, however, there's no such lemma or inflected form in Tibetan. It is most likely meant to be the lemma བསྒྲིག། *zhegh* "arrange" and imperative form སྒྲིགས། *zhi* "arrange". In this case, both the lemma and the inflected form are wrong. This type of error is quite common and amplifies the error rates.

5.2 Prediction errors

Table 2 presents the distribution of the number of prediction errors for each system trained in medium-resource setting. Below we present a more detailed analysis of each error category.

5.2.1 Nonce-word errors

This type of errors corresponds to illegal words, i.e. situations when the string generated by a system does not exist in Tibetan. We identify two sub-types of nonce-word errors in the Tibetan language.

The first one occurs because the Tibetan script is 2-dimensional (see Section 2),

[5] The hyperparameters are set to the values reported in the corresponding papers.

[6] Most language data in the UniMorph dataset was automatically extracted from Wiktionary.

whereby affixes may appear in a total of six positions relative to the base word. As can be seen in the following output:

(1) འབོག *wugh* “cross water” + FUT → ོག (nonce-word)

The correct answer should be འབོགས *wi* “will be crossing water”, whereas the system predicted the suffix of the word but ignored the prefix and the second suffix. Since 2-dimensional scripts such as Tibetan are rare in the world’s writing systems, researchers often do not consider this problem.

The second type relates to special cases in Tibetan. Some components are rarely used, and are unique variants of certain consonants as affixes. This often causes problems for learners of the Tibetan language. The following is an example of this case:

(2) དྭངས *ngi* “clear” + PRS → དཔངས *bi* (nonce-word)

Here, a special component representing the vowel ཝ *wha* should occur under the base component as an affix, meaning the correct answer is དྭངས *ngi* “clear”.

5.2.2 Allomorphy errors

Allomorphy errors occur more often than nonce-word ones, and here we also classified them into two sub-types.

Firstly, the rules of Tibetan verb inflection are very complicated and irregular. For some of them, it is impossible for a system to learn the relevant rules through generalization over the training set. For instance, the present tense of the verb ཤི *hi* “die” is not predictable from its lemma འཆི *qie* “die”. In this experiment, as can be seen in the example below, where the correct output should be ཟོ *su* “eat”, the system attempts to inflect based on rules that it has learned which are inappropriate for this word:

(3) བཟའ *sa* “eat” + PRS→ ཚུས *tsi* (nonce-word)

The second error type relates to vowels where the systems fail to predict the correct position of diacritics. Diacritics are vowels and can only be added above or below the base components, but systems fail to learn this constraint, and over-generate diacritic positions. This kind of error also occurs when the models add an affix to the wrong position, or the order of affixes is incorrect, i.e. the models have predicted the components correctly but are unable to predict the correct order, as occurs in the following case (where the correct output is ཟད *dep* “exhaust”):

Model	Low	Med	High
Lemma Copy	0.70	0.70	0.70
SMP Baseline	0.65	0.61	0.61
A&G 2017	$0.19^{.05}$	$0.52^{.03}$	$\mathbf{0.85}^{.02}$
M&C 2018	$\mathbf{0.73}^{.03}$	$\mathbf{0.72}^{.02}$	$0.76^{.02}$

Table 3: Accuracy for the systems trained using the corrected dataset (best in bold).

(4) ཟད *dep* “exhaust” + IMP →དཟ *wegh* (nonce-word)

5.3 Results obtained on a new dataset

Since we found many target errors in the UniMorph data, we used the new verbal lexicon to improve the linguistic fidelity of the setup.

We manually counted 103 target errors out of 158 samples in the medium training set, which is 65% of the dataset. After correcting all the target errors in all sets according to the new lexicon, we reran our experiments.

Since both A&G 2017 and M&C 2018 use random parameter initialization, we ran the models with five different random seeds, and report their mean accuracy along with standard deviation. As Table 3 shows, in all three settings and across all three trained systems, the best accuracy increases substantially. Table 4 also provides the distribution of the number of prediction errors made on the corrected data. While there is clearly substantial room for improvement in the results, we believe these results to be much more reflective of the true ability of contemporary morphological reinflection systems to model Tibetan.

6 Conclusion

We focused on Tibetan verbal morphology in the context of sub-task-1 of the SIGMORPHON 2018 shared task. We considered a range of baselines and two state-of-the-art models trained in different data size

Model	Nonce	Allomorphy
Lemma Copy	0	15
SMP Baseline	5	16
A&G 2017	3	8
M&C 2018	8	14

Table 4: Absolute number of errors on the test set made by each system trained in high-resource setting (corrected data).

conditions. After conducting a detailed error analysis, we discovered that a significant percentage of errors relate to noise of the data and irregularity of Tibetan. We re-annotated the data and also developed a new Tibetan verbal lexicon comprising 1,433 lemmata with corresponding inflected forms. After re-running the model on the clean data, we observed a substantial improvement in terms of accuracy.

A possible research direction for future work would be to tailor the models to the idiosyncrasies of the Tibetan language.

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2004–2015, Vancouver, Canada.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection, pages 1–27, Brussels, Belgium.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection, pages 1–30, Vancouver, Canada.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but OK: Making sense of morphological generation errors. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 140–151, Hong Kong.

Nathan W Hill. 2010. A lexicon of Tibetan verb stems as reported by the grammatical tradition. Bayerische Akademie der Wissenschaften.

Jitaiga. 2018. The progress of Tibetan natural language processing. 34(1):1–11. Forum of Tibetan Development.

Jitaijia. 2013. Research on Tibetan Syntax.

Gesang Jumian. 1982. The category of Tibetan verbs. Minority Language of China, (5):27–39.

Yonghong Li, Yixin Zhou, Jing Shi, and Hongzhi Yu. 2009. On the origin of Tibetan language. Journal of Language and Literature Studies, (3):31–34.

Peter Makarov and Simon Clematide. 2018. Neural transition-based string transduction for limited-resource setting in morphology. In Proceedings of the 27th International Conference on Computational Linguistics, pages 83–93, Santa Fe, USA.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. arXiv preprint arXiv:1701.03980.

Aitang Qu. 1985. The structure and evolution of the inflectional forms of Tibetan verbs. Minority Language of China, (1):1–15.

Aitang Qu and Song Jing. 2011. The 'theory of sound' and the principle of Tibetan creation. Minority Language of China, (5):15–25.

Suonanjiancuo. 2013. Study on the adhesion and inflection of Tibetan verbs. Journal of University of Tibet, (4):70–75.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International

Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 674–680, Beijing, China.

Yao Wang. 1980. A brief account of Tibetan ancient historical documents. Journal of Xizang Minzu University, (2):11–37.

Cai Zhijie. 2005. Research on the development of Tibetan–Chinese–English electronic dictionary. Journal of Qinghai Norma University, (2):48–50.

Cai Zhijie. 2009. The design and realization of Tibetan spelling. Journal of Qinghai Normal University, (1):69–71.

Cai Zhijie and Cai Rangzhuoma. 2010. Design of BanZhiDa Tibetan lexicon. Journal of Chinese Information Processing, 24(5):46–50.