SimpleNLG-TI: Adapting SimpleNLG to Tibetan

Zewang Kuanzhuo
Tibetan information processing and Machine Translation
Key Laboratory of Qinghai Province, Qinghai Normal University
khagro@qq.com
Lin Li
Information and Computing Sciences, Utrecht University
Department of Computer Science, Qinghai Normal University
l.li1@uu.nl
Weina Zhao
Department of Computer Science, Qinghai Normal University
490333294@qq.com

Abstract

Surface realisation is the last but not the least phase of Natural Language Generation, which aims to produce high-quality natural language text based on meaning representations. In this article, we present our work on SimpleNLG-TI, a Tibetan surface realiser, which follows the design paradigm of SimpleNLG-EN. SimpleNLG-TI is built up by our investigation of the core features of Tibetan morphology and syntax. Through this work, we provide a robust and flexible surface realiser for Tibetan generation systems.

1 Introduction

Natural Language Generation(NLG) aims to produces Natural Language based on various input like meaning representations or numeric data. NLG systems can be classified into text-to-text or data-to-text system according to the system input. Most early generation work in Tibetan focused on text-to-text generation by Neural Networks, for instance, Poem Generation(Chajia, 2019), Machine Translation(Yachao, 2017), Question Answering(Sun et al., 2019; Xiaoying, 2017). There is still a lack of theoretical studies and available resources for Tibetan generation research. To solve the latter problem, this paper offers a practical tool, SimpleNLG-TI, for generating Tibetan text.

Classical pipeline model(Reiter and Dale, 1997) proposed that a generation process shall accomplish six tasks, according to which the last but not the least phase of the process is surface realisation. Surface realisation aims to produce strings that conform to grammar rules, therefore it plays an important role on improving readability of output. To build up SimpleNLG-TI, we solve key theoretical and practical problems in Tibetan realising. Firstly, we build up a linguistic rule bank for SimpleNLG-TI by exploring morphology and syntax features in Tibetan such as function words, inflection, word order. Secondly, in order to develop a practical surface realiser we investigate how to apply the framework of SimpleNLG-EN to SimpleNLG-TI properly and reuse its code[1].

The rest of this paper is organised as follows. In section 2, we discuss some related work. In section 3, we discuss Tibetan morpho-syntactic characteristics and how to apply them into SimpleNLG-TI. Evaluation process is introduced in section 4. And we end with conclusion and future work.

2 Related Work

As an important member of Sino-Tibetan language, Tibetan has 3 major dialect areas in China and is used by about 8 million people around the world. A classic work(de Körös, 1834) analyzed phonetic and morphology, and syntax features of Tibetan. A primary feature of Tibetan is its SOV word-order, for instance, in the predicate of example (1) ཟ (*eat*) locates at the end of the sentence. Moreover, there are abundant reductions in Tibetan. Both the original case ཀྱིས and its reduction form ས are correct in example (1). In most situation, the reduction form is preferred.

[1]We referred to SimpleNLG v4.4.8 rewrite by Bjascob https://github.com/bjascob/pySimpleNLG.

Proceedings of The 13th International Conference on Natural Language Generation, pages 86–90,
Dublin, Ireland, 15-18 December, 2020.

(1) Tibetan: ངས་ཤ་ཟ། (= ང་ཡིས་ཤ་ཟ།)
English: *I eat meat.*

In this work, the coverage of SimpleNLG-TI mainly refer to two highly-respected grammar books(Jumian, 1987; de Körös, 1834), which offer us clear and operable grammar. Besides of the earlier work of linguistics, rich language resources have been collected for studies in data-driven Tibetan processing, for instance, a large scale Tibetan corpus has been built up (Liu et al., 2012) which provides SimpleNLG-TI with an available vocabulary.

As a significant phase of NLG, surface realisation has attracted much research attention and many surface realisers have been developed such as KPM(Bateman, 1997), YAG(McRoy et al., 2000), and OPENCCG(White, 2006). SimpleNLG-EN is an open source project(Gatt and Reiter, 2009,?), which allows it to be integrated into various language generation applications. For instance, in sentence generating(Sheikha and Inkpen, 2011) and sentence summarizing(Khan et al., 2015). To realise well-formed text, SimpleNLG-EN adopts rule-based approach based on lexicon and rules with rich linguistic knowledge. The paradigm of SimpleNLG has been successfully implemented into many natural language such as Chinese SimpleNLG-ZH(Chen et al., 2018) and German SimpleNLG-DE(Braun et al., 2019), and SimpleNLG-EnFr(Vaudry and Lapalme, 2013).

3 Coverage of SimpleNLG-TI

SimpleNLG-TI employs the framework of SimpleNLG-EN because of its comprehensive linguistic component definition. In process of designing and developing SimpleNLG-TI, we firstly build up a component set contains content words, function words, and highly-frequent phrases. Then, to realise linguistic components into a well-formed string, we also establish a rule bank includes orthography, morphology and syntax rules by making use of linguistic knowledge in Tibetan grammar books.

The rich morphological features of Tibetan lead to challenge for surface realisation task. In this work, we only focus on core morphological and syntactic characteristics that play important roles on realisation. Table 1 shows how generates the following sentence by SimpleNLG-TI.
Tibetan: ཤིང་མཁན་གྱིས་ཤིང་ལྗུ་རེས་ཡིས་བཅད།
English: *A carpenter cut down a tree with axe.*

Input Code
lexicon = XMLLexicon()
nlgFactory = NLGFactory(lexicon)
sentence = nlgFactory.createClause()
sentence.setSubject("ཤིང་མཁན་")
sentence.addCase("གྱིས་")
sentence.addComplement("ལྗུ་རེས་ཡིས་")
sentence.setVerb("བཅད་")
output = Realiser.realiseSentence(sentence)

Table 1: An instance of realising a sentence with SimpleNLG-TI.

3.1 Morphology

Nouns and verbs are morphology varied in Tibetan, therefore SimpleNLG-TI takes inflections of nouns and verbs into account because of its significant role in surface realising.

3.1.1 Collective markers

In Tibetan, the concept of plural is realised by using various plural marks like རྣམས་, ཚོ་, ཅག, དག, སཁའ་དག, ཀུན, ཐམས་ཅད, ཡོངས, and སོགས. SimpleNLG-TI can realise four commonly used plural marks, i.e., རྣམས་, ཅག, ཚོ་, and དག. The first three marks can be used with most nouns, yet the last one usually only following inanimate nouns. We provide several examples of different usages of plural marks as followed.

Example (2) shows a common noun *student* with tow different plural marks རྣམས and དག respectively, both of which express the meaning of students. An example of inanimate noun *table* is made in example (3). example (4) and (5) are instances of how to expressing plural concept of pronouns by ཚོ and ཅག.

(2) སློབ་མ(*student*) སློབ་མ་རྣམས(*students*)
སློབ་མ(*student*) སློབ་མ་དག(*students*)

(3) སདུན་ཅོག(*table*) སདུན་ཅོག་དག(*tables*)

(4) ང(*I*) ང་ཚོ(*we*)

(5) ཁོ(*he*) ཁོ་ཅག(*they*)

3.1.2 Verb Tense

Tibetan verbs have three forms of inflection, that is tense, imperative and the change of the voices. Clearly rules of morphology changes are not founded yet, therefore SimpleNLG-TI has adopted a simplified model for verb realising. example (6) is the three forms of verb ཟ *eat*. example (7) is a special case whose three inflection are the same, which is not uncommon in Tibetan. The lexicon of SimpleNLG-TI contains more than 1,500 monosyllabic verbs for now, which includes the above inflection information.

(6) ཟ བཟས བཟའ ཤོ
eat *ate* *will eat* *do eat*
present *past* *future* *imperative*

(7) རེག རེག རེག རེག
touch *touched* *will touch* *do touch*
present *past* *future* *imperative*

3.2 Syntax

A simple sentence is normally composed of nouns (or noun phrases), verbs (or verb phrases), and function words such as case and aspect. In this section, we discuss word order in Tibetan briefly, then explain how SimpleNLG-TI realises noun and verb phrases.

3.2.1 Word order

Primary word order in Tibetan is Subject-Object-Verb(SOV), which means predicate normally locates at the end of a sentence. A predicate is normally composed of a verb (or verb phrase) or an adjective (or adjective phrase)(Li and Long, 2012).

Example (8) is a simple sentence which is composed of a noun and verb phrase. Take the process of realising (8) as an example, SimpleNLG-TI firstly realises the noun as an independent component, and then separately generates the elements of the verb phrase as an object and complement. The sentence mentioned earlier in this section can be realised by SimpleNLG-TI as sentence (9). In addition to realising a noun phrase and a verb phrase in the same way of (8), the case mark གྱིས is an essential input.

(8) Tibetan: ངའི་རྒྱུད་ཡིན།
English: *I teacher* ***am***

(9) Tibetan: ཤིང་མཁན་གྱིས་ཤིང་ལྟེ་རེས་ཡིས་བཅད།
English: *The carpenter the tree with an axe* ***cut down***.

3.2.2 Negation

Negation words are མ, མི, མེད, and མིན, which modify nouns, verbs and adjectives. In this work, we follow investigations proposed by(Jieben, 2012), thus the rules of negation words using by SimpleNLG-TI are listed here.

1. Both མ and མི are treated as prepositional negatives, which are used to modify verbs and adjectives.

2. Besides of POS (i.e., part-of-speech), tense also determines the choice of negation words. In SimpleNLG-TI, མི is usually used in present and future tense like (10) and མ often used in past tense such as example (11).

(10) སློབ་གྲྭ་འགྲོ་ སློབ་གྲྭ་མི་འགྲོ་
go to school *don' t go to school*
ཨ་འོངས་པར་སློབ་གྲྭ་འགྲོ་ ཨ་འོངས་པར་སློབ་གྲྭ་ལ་མི་འགྲོ་
will go to school *will not go to school*

(11) སློབ་གྲྭར་སོང་། སློབ་གྲྭར་མ་སོང་།
went to school *didn't go to school*

3. མེད and མིན is negation of ཡོད and ཡིན.

(12) ངའི་དགེ་རྒན་ཡིན། ངའི་དགེ་རྒན་མིན།
I'm a teacher. *I'm not a teacher.*
ངལ་དེའི་ཆ་ཡོད། ངལ་དེའི་ཆ་མེད།
I have book *I don' t have book.*

3.2.3 Noun phrase and Adjective phrase

SimpleNLG-TI is capable of realising a noun phrase whose head word is modified by an adjective or noun. When modification is an adjective, SimpleNLG-TI can realised a noun phrase in two ways:

1. noun + noun/adjective like example (13)
2. adjective/pronoun + case + noun such as example (14)

The realisation of an adjective phrase follows the way of noun phrases, for instance, example (15). The adjective phrase is realized as a noun phrase, the adjective is the center word which is modified by the adverb.

(13) མེ་ཏོག་ + མཛེས་པ་ = མེ་ཏོག་མཛེས་པ་
flower + *beautiful* = *flower beautiful*
རྟ་ + ཁྱོ་ + ལུག་ = རྟ་ཁྱོ་ལུག་
horse + *yak* + *sheep* = horse yak sheep

(14) མཛེས་པ་ + དེ་ + མེ་ཏོག་ = མཛེས་པའི་མེ་ཏོག་
beauty + case + flower = beautiful flower

(15) ཤིན་ཏུ་ + མཛེས་ = ཤིན་ཏུ་མཛེས་
very + beautiful = very beautiful

3.2.4 Verb phrase

Structure of verb phrases is varied, which is partial because the head words can be simultaneously modified by various components such as auxiliary, adverb, etc.

According to whether a verb phrase contains case or not, SimpleNLG-TI realises verb phrases as follows.
1. noun + verb: in example (16) and (17)
2. noun + (case) + verb: object-verb structure realising such as example (18) and (19)

(16) ཡི་གེ་ བྲིས་ ཡི་གེ་བྲིས་
word wrote wrote word

(17) ཟ་མ་ ཟ་ ཟ་མ་ཟ་
eat food eat food

(18) ཤར་ཕྱོགས་ སུ་ སྐྱོད་ ཤར་ཕྱོགས་སུ་སྐྱོད་
east case go go to east

(19) མཇུག་ལོ་ རུ་ འགྲོ་ མཇུག་ལོངས་འགྲོ་
quickly case go go quickly

3.3 Punctuation

SimpleNLG-TI can realise two punctuation, that is vertical symbol ། and phonetic node ་. As one of the most widely-used punctuation in Tibetan, phonetic node ་ is used to separate two syllables and vertical symbol ། to indicate boundary of two words, pause of words and sentences, and full stop of sentences. In some cases, there are exception in the usage of them.

When a sentence is ended with the character ག, the sentence ends without vertical symbol ། such as example (10). If the last letter of a sentence is ང, the sentence ends with the combination of ་ and ། as shown in example (11).

Normally, the phonetic node ་ is used similar with the white space in English such as example (13). In the situation of reduction raised by case, the phonetic node ་ will be removed. For instance, in example (14) phonetic node ་ of the syllable པ་ is removed the case དེ is reduced.

4 Evaluation

Following the evaluation approach in early work(Soto et al., 2017; Fuentes et al., 2018), SimpleNLG-TI is evaluated by comparing its output with a golden standard. When an output is completely identical with the golden standard, then we take SimpleNLG-TI passes this unit test.

In this work, we take a test set as our golden standard, which is manually translated from the test set of SimpleNLG-EN. SimpleNLG-TI has successfully passed all 84 unit tests, which cover Tibetan linguistic features previous described in this paper. The result shows that this work can be used as a practical tool for NLG in Tibetan.

5 Conclusion

In this work, we propose SimpleNLG-TI (a Surface Realiser) for Tibetan generation, which is based on the framework of SimpleNLG-EN. We investigate main linguistic features (morphology and syntactic characteristics) and apply them into the development of SimpleNLG-TI. The realisation results of SimpleNLG-TI show that it is capable of generating well-formed text and is feasible to be deployed in Tibetan generation systems.

In the future, we plan to improve the performance of SimpleNLG-TI by taking more features into account, for instance, other types of phrases like adjective phrases and complex words and phrases. Moveover, we would like to test SimpleNLG-TI with practical Tibetan generation systems. The full Python package of SimpleNLG-TI will be publicly released on Github[2].

Acknowledgments

The authors of this paper received support from Qinghai Natural Science Foundation underGrant 2016-ZJ-931Q, Qinghai Major RD Transformation Foundation under Grant 2019-GX-162, and National Natural Foundation under Grant 61862055, which is gratefully acknowledged.

[2] https://github.com/ZWKZ/SimpleNLG-TI

References

John A Bateman. 1997. Enabling technology for multilingual natural language generation: the kpml development environment. *Natural Language Engineering*, 3(1):15–55.

Daniel Braun, Kira Klimt, Daniela Schneider, and Florian Matthes. 2019. Simplenlg-de: Adapting simplenlg 4 to german. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 415–420.

Se Chajia. 2019. *Tibetan Poem Generation with Attention Based Encoder-Decoder Model*, volume 33. Journal of Chinese Information Processing.

Guanyi Chen, Kees Van Deemter, and Chenghua Lin. 2018. Simplenlg-zh: A linguistic realisation engine for mandarin. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 57–66.

Andrea Cascallar Fuentes, Alejandro Ramos Soto, and Alberto Bugarin Diz. 2018. Adapting simplenlg to galician language. *Proceedings of the 11th International Conference on Natural Language Generation*, pages 67–72.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93.

Dao Jieben. 2012. *The Grammatical Functions of the Tibetan Adverbs*. Ph.D. thesis, Northwest University for Nationalities.

Gesang Jumian. 1987. Practical tibetan grammar. *Sichuan Minorities Press*.

Atif Khan, Naomie Salim, and Yogan Jaya Kumar. 2015. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30:737–747.

Sandor Csoma de Körös. 1834. *A grammar of the Tibetan language*. Baptist Mission Press.

Lin Li and Congjun Long. 2012. Recognition of tibetan linking verb and existential verb. In *Journal of Chinese Information Processing*.

Huidan Liu, Minghua Nuo, Jian Wu, and Yeping He. 2012. Building large scale text corpus for tibetan natural language processing by extracting text from web. In *24th International Conference on Computational Linguistics*, page 11.

Susan W McRoy, Songsak Channarukul, and Syed S Ali. 2000. Yag: A template-based generator for real-time systems. In *Proceedings of the first international conference on Natural language generation-Volume 14*, pages 264–267. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 187–193. Association for Computational Linguistics.

Alejandro Ramos Soto, Julio Janeiro Gallardo, and Alberto Bugarin Diz. 2017. Adapting simplenlg to spanish. *Proceedings of the 10th International Conference on Natural Language Generation*.

Yuan Sun, Chaofan Chen, Tianci Xia, and Xiaobing Zhao. 2019. Qugan: Quasi generative adversarial network for tibetan question answering corpus generation. *IEEE Access*, 7:116247–116255.

Pierre-Luc Vaudry and Guy Lapalme. 2013. Adapting simplenlg for bilingual english-french realisation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 183–187.

Michael White. 2006. Ccg chart realization from disjunctive inputs. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 12–19.

Chen Xiaoying. 2017. Design and research of tibetan encyclopedia knowledge question answering system. *Intelligent Computer and Application*, 7(4):48–50.

Li Yachao. 2017. Research on tibetan-chinese neural machine translation. *Journal of Chinese Information Processing*, 31(6):103–109.