A rule based tagger for Classical Tibetan: negation and verb stems classification

Nathan W. Hill and Edward Garrett

(SOAS, University of London)
Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger
Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Here we focus on the development of rule-based part-of-speech tagging.
Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Here we focus on the development of rule-based part-of-speech tagging.

1. An overview of the project workflow
Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Here we focus on the development of rule-based part-of-speech tagging.

1. An overview of the project workflow
2. A simple example of a rule used by the tagger (so ‘tooth’ or verb suffix)
Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Here we focus on the development of rule-based part-of-speech tagging.

1. An overview of the project workflow
2. A simple example of a rule used by the tagger (so 'tooth' or verb suffix)
3. Distinguishing *mi* and *ma* as nouns and negation markers
Tibetan in Digital Communication

Goals

1. 1,000,000 word part-of-speech tagged corpus of Tibetan texts
2. An automatic word breaker
3. An automatic part-of-speech tagger

Here we focus on the development of rule-based part-of-speech tagging.

1. An overview of the project workflow
2. A simple example of a rule used by the tagger (so 'tooth' or verb suffix)
3. Distinguishing mi and ma as nouns and negation markers
4. Disambiguating verb stems (main focus of the presentation)
### Workflow:

(1) Look-up of possible analyses

<table>
<thead>
<tr>
<th>Word</th>
<th>Part-of-speech tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>སྐད་ཐོབ་</td>
<td>n.count</td>
</tr>
<tr>
<td>་</td>
<td>d.dem ~ cv.sem</td>
</tr>
<tr>
<td>ར་</td>
<td>case.all ~ n.count</td>
</tr>
<tr>
<td>བཀྲ་ཤིས་</td>
<td>n.count</td>
</tr>
<tr>
<td>རྟོག་</td>
<td>num.card</td>
</tr>
<tr>
<td>སྦྱར་</td>
<td>num.card</td>
</tr>
<tr>
<td>དུ་་</td>
<td>v.invar</td>
</tr>
<tr>
<td>བ།</td>
<td>cl.focus</td>
</tr>
<tr>
<td></td>
<td>punc</td>
</tr>
</tbody>
</table>
# Workflow:
(2) Pre-tagging

<table>
<thead>
<tr>
<th>Word</th>
<th>Part-of-speech tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>བུད།</td>
<td>n.count</td>
</tr>
<tr>
<td>བླ་</td>
<td>d.dem</td>
</tr>
<tr>
<td>ལ།</td>
<td>case.all ~ n.count</td>
</tr>
<tr>
<td>ཧི་</td>
<td>n.count</td>
</tr>
<tr>
<td>འདོད།</td>
<td>num.card</td>
</tr>
<tr>
<td>གཉིས་</td>
<td>num.card</td>
</tr>
<tr>
<td>བདེབས།</td>
<td>v.invar</td>
</tr>
<tr>
<td>དུས་</td>
<td>cl.focus</td>
</tr>
<tr>
<td>།</td>
<td>punc</td>
</tr>
</tbody>
</table>
### Workflow:
**3) Hand-tagging**

<table>
<thead>
<tr>
<th>Word</th>
<th>Part-of-speech tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>ཆགས་པ་</td>
<td>n.count</td>
</tr>
<tr>
<td>དེ་</td>
<td>d.dem</td>
</tr>
<tr>
<td>བ</td>
<td>case.all</td>
</tr>
<tr>
<td>སིང་</td>
<td>n.count</td>
</tr>
<tr>
<td>བ</td>
<td>num.card</td>
</tr>
<tr>
<td>བ</td>
<td>num.card</td>
</tr>
<tr>
<td>དོན་</td>
<td>v.invar</td>
</tr>
<tr>
<td>ས་</td>
<td>cl.focus</td>
</tr>
<tr>
<td>བ</td>
<td>punc</td>
</tr>
</tbody>
</table>
Workflow:
(4) Rule suggestions

Screen shot of rule suggestions
(9 November 2013)

- case.ela ← cv.ela
- case.gen ← cv.gen
- n.count ← case.term
- n.v.fut.n.v.pres ← n.v.pres
- n.v.fut ← n.v.fut.n.v.past
- n.v.invar ← n.v.past
- n.v.past.n.v.pres ← n.v.pres
- neg ← n.count
- v.fut.v.pres ← v.invar
- v.invar ← dunno
- v.invar ← v.fut.v.pres
Workflow:
(4) Rule suggestions

Screen shot of the rule suggestion [neg] ← [n.count] (9 November 2013)
A simple example: Identifying sandhi determined suffixes

so 'tooth' versus so finite verb suffix
A simple example: 
Identifying sandhi determined suffixes

so 'tooth' versus so finite verb suffix
finite verb suffix always after -s, but
A simple example:
Identifying sandhi determined suffixes

so 'tooth' versus so finite verb suffix
finite verb suffix always after -s, but

_khos so bcag_'he broke teeth'
A simple example: Identifying sandhi determined suffixes

so 'tooth' versus so finite verb suffix

finite verb suffix always after -s, but

\textit{khos so bcag} 'he broke teeth'

will yield false positive
A simple example:
Identifying sandhi determined suffixes

so 'tooth' versus so finite verb suffix

finite verb suffix always after -s, but

khos so bcag 'he broke teeth'

will yield false positive

So, we also specify that after so is ǵad, or the syllables žes, sñam, or zer.
A simple example:
Identifying sandhi determined suffixes

so 'tooth' versus so finite verb suffix

finite verb suffix always after -s, but

\textit{khos so bcag} 'he broke teeth'

will yield false positive

So, we also specify that after so is \textit{sad}, or the syllables \textit{̱es}, \textit{ṣnam}, or \textit{zer}.

Similar rules for -\textit{lo}, -\textit{ro}, -\textit{de}, -\textit{śiṅ}, etc.
Disambiguating *mi* and *ma* as negations or nouns

(1). Finding the nouns *mi* and *ma* within noun phrases

*mi ḥdi 'this person'*

*mi gsum 'three people'*

*ma rgan-mo 'your old mother'*
Disambiguating *mi* and *ma* as negations or nouns

(2). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

\[ rmo\text{-}pa \ \hat{i} \ \text{mi} \ \hat{ho} \ \text{gro} \ \hat{ho} \]

'an ignorant person goes'.

Disambiguating *mi* and *ma* as negations or nouns

(2). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

*rmoṅ-pa ḥi mi ḥgro ḥo*
'An ignorant person goes'.

*bskal-pa graṅs med-pa ḥi mi dge-ba ḥi las*
'Non virtuous deeds of countless eons'.
Disambiguating *mi* and *ma* as negations or nouns

(2). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

*rmoṅ-pa ḥi mi ḥgro ḥo*

'an ignorant person goes'.

*bskal-pa graṅs med-pa ḥi mi dge-ba ḥi las*

'non virtuous deeds of countless eons'.

*rab tu ḥbyuṅ-ba ḥi mi rigs*

'it is not proper to take ordination'.
Disambiguating *mi* and *ma* as negations or nouns

(2). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

*rmoṅ-pa ḥi mi ṭgro ḥo*

'an ignorant person goes'.

*bskal-pa graṅs med-pa ḥi mi dge-ba ḥi las*

'non virtuous deeds of countless eons'.

*rab tu ḥbyuṅ-ba ḥi mi rigs*

'it is not proper to take ordination'.

RULE: If *mi* / *ma* could be [n.count], follows a probable genitive, does not precede *rigs*, and does not precede a [n.v.xxx], and the word before the probable genitive is not an unambiguous [v.xxx] tag, then mark *mi* / *ma* as a [n.count].
Disambiguating *mi* and *ma* as negations or nouns

(2). Isolating *mi* [n.count] and *ma* [n.count] after the genitive

*rmol-pa Ḣi mi Ḣgro Ḥo*

'an ignorant person goes'.

*bskal-pa graṅs med-pa Ḣi mi dge-ba Ḣi las*

'non virtuous deeds of countless eons'.

*rab tu Ḥbyuṅ-ba Ḣi mi rigs*

'it is not proper to take ordination'.

RULE: If *mi* / *ma* could be [n.count], follows a probable genitive, does not precede *rigs*, and does not precede a [n.v.xxx], and the word before the probable genitive is not an unambiguous [v.xxx] tag, then mark *mi* / *ma* as a [n.count].

PATTERN: \($S + \| (?:\[(?!v.)[^\]]*)\) + \s + (?:\ल | ल | ल) | \S + \s + (?:\स | स) | \S + \S*\[n\.count\]* | \S* (?!\s + (?:\ङ्ग | ङ्ग) \| \S + \[n\.v\].)\)

REPLACE: $1|[n.count]
Disambiguating *mi* and *ma* as negations or nouns

(3). Isolating *mi* and *ma* as nouns before *med-pa*

*mi med-paḥi sa* 'a place without people'

*bu ma med-pa* 'a boy without a mother'
Disambiguating *mi* and *ma* as negations or nouns

(4). Using the associative case to find nouns

*lha dan mi* 'gods and men'
Disambiguating *mi* and *ma* as negations or nouns

(5). Identifying *mi* [neg] before present and future tense verb forms

*mi byed do* 'doesn't do'

*mi gšegs so* 'doesn't go'
Disambiguating *mi* and *ma* as negations or nouns

(5). Identifying *mi* [neg] before present and future tense verb forms

*mi byed do* 'doesn't do'

*mi gsëgs so* 'doesn't go'

*mi gari* 'not full' and 'which person'
Disambiguating *mi* and *ma* as negations or nouns

(5). Identifying *mi* [neg] before present and future tense verb forms

- *mi byed do* 'doesn't do'
- *mi gšegs so* 'doesn't go'
- *mi garî* 'not full' and 'which person'
- *mi sogś te* 'people etc’
Disambiguating *mi* and *ma* as negations or nouns

(5). Identifying *mi* [neg] before present and future tense verb forms

*mi byed do 'doesn't do'*

*mi gšëgs so 'doesn't go'*

*mi gañi 'not full' and 'which person'*

*mi sogs te 'people etc'*

*mi bgres-pa 'aged person'
Disambiguating *mi* and *ma* as negations or nouns

(6). Identifying *ma* [neg] in the prohibitive

*ma byed* 'do not do it!'

*ma gšegs sig* 'do not go!'
Disambiguating *mi* and *ma* as negations or nouns

(7). Isolating *ma* [neg] before the past tense and *yin*

*ma byas* 'didn't do it'

*ma gṣegs* so 'didn't go'

*ma yin te* 'isn't' (rarely 'is a mother')
Distinguishing the four tenses and subsequent cleanup

As an default hypothesis, we follow the dictionaries (cf. Hill 2010) in assuming that all Tibetan verbs in principle distinguish four tenses, the present [v.pres], the past [v.past], the future [v.fut], and the imperative [v.imp].

Disambiguating verb tenses

1. Morphology
2. Syntax
3. \textit{da-drag} sandhi
Morphology

No problem when all four tenses are distinct.

pres. byed, past byas, fut. bya, imp. byos 'do'

pres. gsod, past bsad, fut. gsad, imp. sod 'kill'
Morphology

No problem when all four tenses are distinct.

pres. byed, past byas, fut. bya, imp. byos 'do'

pres. gsod, past bsad, fut. gsad, imp. sod 'kill'

A problem when all four stems are not distinct.

pres. gšegs, past gšegs, fut. gšegs, imp. gšegs 'go'

pres. mdzad, past mdzad, fut. mdzad, imp. mdzod 'do'

pres. skyel, past bskyal, fut. bskyal, imp. skyol 'carry'
Syntax: Using verbal suffixes

(1). Finding the imperative before [cv.imp]

*byos šig 'do it!' (imperative)*

*gšegš šig 'go!' (imperative)*
Syntax:
Using verbal suffixes

(1). Finding the imperative before [cv.imp]

byos śig 'do it!' (imperative)

gśegs śig 'go!' (imperative)

But watch out

ma gśegs śig 'do not go!' (present)

gnas śig 'stay!' (imperative) or 'a place' (noun)
Syntax: Using verbal suffixes

(2). Finding the prohibitive (present negated with *ma*) before [cv.imp]

*ma byed cig 'do not do!' (present)*

*ma gšegs šig 'do not go!' (present)*
(3). Prohibiting the imperative in non-finite contexts

(None of the following are the imperative.)

- gśegs na 'if he goes'
- gśegs kyaṅ 'although he goes'
- gśegs nas 'after going'
- gśegs śîṅ 'goes and ... '
- gśegs te 'going... ' 
(3b). Prohibiting the imperative in explicitly non-imperative, but finite contexts

(Neither of the following is an imperative.)

\textit{g̟egs sam} 'is he going?'

\textit{g̟egs so} 'he goes'

Syntax:  
Using verbal suffixes
Syntax: Using verbal suffixes

(4). The prohibition of the future before the elative converb nas

*byed nas 'after doing' (present)

byas nas 'after doing' (past)

*bya nas (unattested)

gșegs nas is not future (present or past)
(5). Negation with *ma* implies past

*ma byas so* 'didn't do' (past)

*ma gšegs so* 'didn't go' (past)
(6). Negation with *mi* precludes past

*mi byed* 'isn't doing' (present)

*mi bya* 'won't do' (future)

*mi gšegs* 'isn't going, won't go' (present or future)

(N.B. *mi gšegs nas* present only)
**da-drag sandhi**

<table>
<thead>
<tr>
<th>Future</th>
<th>Past</th>
</tr>
</thead>
<tbody>
<tr>
<td>bskyal lo</td>
<td>bskyal to</td>
</tr>
<tr>
<td>bskyal lam</td>
<td>bskyal tam</td>
</tr>
<tr>
<td>bskyal du</td>
<td>bskyal tu</td>
</tr>
<tr>
<td>bskyal yaṅ</td>
<td>bskyal kyaṅ</td>
</tr>
<tr>
<td>bskyal ba</td>
<td>bskyal pa</td>
</tr>
</tbody>
</table>
da-drag sandhi

Not all past forms have a da-drag.

'give' (sbyin, byin, sbyin, byin), always byin no (not byin to).

Some da-drag forms are presents.

'take' (ḥdzin, bzuni, gzuni, zuvis), always ḡdzin to (not ḡdzin no).
da-drag sandhi

Not all past forms have a da-drag.

'give' (sbyin, byin, sbyin, byin), always byin no (not byin to).

Some da-drag forms are presents.

'take' (ḥdzin, bzuni, gzuni, zuinis), always ḡdzin to (not ḡdzin no).

Because -pa and -ba are similar looking and frequently confused, this rule may seem to risk introducing errors. However, we think it is best to disambiguate verb stems wherever it is possible to do so. Disambiguating these stems permits the behavior of -pa versus -ba to be more easily explored by future researchers; reason enough to add the rule.
The correct ordering of disambiguation strategies

1.

gan žig śin tu dad-pa ḥi sems kyis chu sñam-pa gan tsam saris-rgyas la mchod dam |
dge ḥdun la phul [v.past] ~ [v.imp] lam | pha-ma la phul [v.past] ~ [v.imp] lam | dbul-phoṅs la byin nas | gcan-gzan la byin na | gsod-nams ḥdi ni bskal-ba du-ma r yan mi zad de |

"If one with a mind of great faith offers handfuls of water to the buddha, or makes offerings to the saṃgha, or makes offerings to one's parents, or gives to the poor, or gives to wild animals, this merit shall not run out for many eons".

(pres. ḥbul, past phul, fut. dbul, imp. phul) 'give'
The correct ordering of disambiguation strategies

1.

gañ ḏi gñin tu ḏad-pañ hī sṃs kyi sñam-pañ goñ tsaṃ saṃs-rgyas la mchod dām ḏal ḍhduṅ la phul [v.past] ∼ [v.imp] laṃ ḏal pha-ma la phul [v.past] ∼ [v.imp] laṃ ḏal dbul-phoṅs la byin nas ḏal gcan-gzan la byin na ḏal gsod-nams ḍhdi ni bskal-ba du-ma ṭa yaṅ mi zad de ḏal

"If one with a mind of great faith offers handfuls of water to the buddha, or makes offerings to the saṃgha, or makes offerings to one's parents, or gives to the poor, or gives to wild animals, this merit shall not run out for many eons".

(pres. ḍbul, past phul, fut. dbul, imp. phul) 'give'

If da-drag rules are ordered before the exclusion of [v.imp] before [cv.ques], then the machine concludes that phul is [v.imp], since the suffix should have been –tam for the past.
The correct ordering of disambiguation strategies

2.

ku-ṣu ḥdi ni ḥbras-bu las skyes-pa ma lags te | chab-mig cig gi naṅ nas rñed-pa s slan-cad ni bdag gis mi rñed de | mi ḥbyor [v.past] ~ [v.pres] to

"This apple was not born from fruit, but I found it from inside a spring, so I cannot find it hereafter. It will not be encountered."
The correct ordering of disambiguation strategies

2.

*ku-šu ḥdi ni ḥbras-bu las skyes-pa ma lags te | chab-mig cig gi naṅ nas rñed-pa s slan-cad ni bdag gis mi rñed de | mi ḥbyor [v.past] ~ [v.pres] to*

"This apple was not born from fruit, but I found it from inside a spring, so I cannot find it hereafter. It will not be encountered."

If *da-drag* rules are ordered before the rule that prevents the past stem after *mi*, then the suffix -*to* would have triggered [v.past].
The correct ordering of disambiguation strategies

2.

ku-ṣu ḥdi ni ḥbras-bu las skyes-pa ma lags te | chab-mig cig gi naṃ nas rñed-pa s slan-
cad ni bdag gis mi rñed de | mi ḥbyor [v.past] ~ [v.pres] to

"This apple was not born from fruit, but I found it from inside a spring, so I
cannot find it hereafter. It will not be encountered."

If da-drag rules are ordered before the rule that prevents the past stem after mi,
then the suffix -to would have triggered [v.past].

We order negation rules before da-drag rules, so ḥbyor is tagged [v.pres].
The correct ordering of disambiguation strategies

3.

gal te sñin nas ma btsal [v.past] ~ [v.fut] lam le-lo žig byas te | …

'if one does not seek wholeheartedly, or is lazy…'.
Preliminary observations on the distribution of *da-drag*

1. *zer* 'say' never gets a *da-drag*

2. *bden* 'be true',
   - as a finite verb always *bden no* (without *da-drag*)
   - as a verbal noun always *bden-pa* (with *da-drag*)
We only go so far

*bdag rab tu dbyur du gsol*
'I request that you give me ordination'.

*bdag la saṅs-rgyas kyi chos bśad du gsol*
'I request that you explain to me the Buddha's dharma'.
We only go so far

`bdag rab tu dbyurṅ du gsol`
'I request that you give me ordination'.

`bdag la saṅs-rgyas kyi chos bṣad du gsol`
'I request that you explain to me the Buddha's dharma'.

One could argue that since `dbyurṅ` is a future, then `bṣad` is also a future.
We only go so far

*b dag rab tu dbyuṅ du gsol*
'I request that you give me ordination'.

*b dag la saṅs-rgyas kyi chos bśad du gsol*
'I request that you explain to me the Buddha's dharma'.

One could argue that since *dbyuṅ* is a future, then *bśad* is also a future.

But this is debatable, and hard for a computer to notice.
Consolidating ambiguous verb forms into ambiguous tags

bskyed [v.fut] ~ [v.past] do [cv.fin]
Consolidating ambiguous verb forms into ambiguous tags

\[ bskyped \ [v.fut] \sim \ [v.past] \ do \ [cv.fin] \]

> \[ bskyped \ [v.fut.v.past] \ do \ [cv.fin] \]
Consolidating ambiguous verb forms into ambiguous tags

bskyed [v.fut] ~ [v.past] do [cv.fin]

> bskyed [v.fut.v.past] do [cv.fin]

gśegs [v.fut] ~ [v.past] ~ [v.pres] so [cv.fin]
Consolidating ambiguous verb forms into ambiguous tags

bskyed [v.fut] ~ [v.past] do [cv.fin]

> bskyed [v.fut.v.past] do [cv.fin]

gśegs [v.fut] ~ [v.past] ~ [v.pres] so [cv.fin]

> gśegs [v.invar] so [cv.fin]
Restoring ambiguity when a single form might belong to two distinct verbs

The preceding rules would make the change

źu [v.fut] ~ [v.past] ~ [v.pres]
Restoring ambiguity when a single form might belong to two distinct verbs

The preceding rules would make the change

\[ \acute{z}u \ [v.\text{fut}] \sim [v.\text{past}] \sim [v.\text{pres}] \]

\[ > \acute{z}u \ [v.\text{invar}] \]
Restoring ambiguity when a single form might belong to two distinct verbs

The preceding rules would make the change

\[ \acute{z}u \ [v.\text{fut}] \sim [v.\text{past}] \sim [v.\text{pres}] \]

\[ > \acute{z}u \ [v.\text{invar}] \]

but \( \acute{z}u \ [v.\text{fut}] \sim [v.\text{pres}] \) belong to the verb 'request'

whereas \( \acute{z}u \ [v.\text{past}] \) belongs to the verb 'melt'
Restoring ambiguity when a single form might belong to two distinct verbs

The preceding rules would make the change

\[ \acute{z}u \ [v.\text{fut}] \sim [v.\text{past}] \sim [v.\text{pres}] \]

\[ > \acute{z}u \ [v.\text{invar}] \]

but \( \acute{z}u \ [v.\text{fut}] \sim [v.\text{pres}] \) belong to the verb 'request' whereas \( \acute{z}u \ [v.\text{past}] \) belongs to the verb 'melt'

So, we change

\( \acute{z}u \ [v.\text{invar}] \)
Restoring ambiguity when a single form might belong to two distinct verbs

The preceding rules would make the change

\[ \acute{z}u \ [v.\text{fut}] \sim [v.\text{past}] \sim [v.\text{pres}] \]

\[ \rightarrow \acute{z}u \ [v.\text{invar}] \]

but \( \acute{z}u \ [v.\text{fut}] \sim [v.\text{pres}] \) belong to the verb 'request'

whereas \( \acute{z}u \ [v.\text{past}] \) belongs to the verb 'melt'

So, we change

\[ \acute{z}u \ [v.\text{invar}] \]

\[ \rightarrow \acute{z}u \ [v.\text{fut}\cdot\text{v.}\text{pres}] \sim [v.\text{past}] \]
Other such cases

za [v.invar] 'itch'

versus za [v.pres] 'eat'

mchis [v.invar] 'be'

versus mchis [v.past] 'go'
### How well does it work?

#### Accuracy and Ambiguity

<table>
<thead>
<tr>
<th></th>
<th>Mdzangs blun</th>
<th></th>
<th>Milarepa namthar</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Ambiguity</td>
<td>Accuracy</td>
<td>Ambiguity</td>
</tr>
<tr>
<td>LexTagger</td>
<td>1.000</td>
<td>2.415</td>
<td>1.000</td>
<td>2.441</td>
</tr>
<tr>
<td>RuleTagger</td>
<td>0.988</td>
<td>1.403</td>
<td>0.991</td>
<td>1.430</td>
</tr>
</tbody>
</table>

*(on 11 Dec, 2013)*
Thank you