The Use of yig-cha and chos-kyi-rnam-grangs in Computing Lexical Cohesion for Tibetan Topic Boundary Detection

Paul G. Hackett
Columbia University
August 18, 2010
Tibetan IT Panel — IATS-12, Vancouver, BC
Introduction

- Simple Tibetan IR system requires segmentation (n-gram, POS-tagging, dictionary substring matching, etc.)
- For finer grain indexing, large-scale structure and (sub-)topic detection is needed
In a previous paper (Hackett, 2000), we reported on automatic techniques developed for Tibetan for:

- Word Segmentation,
- Part-Of-Speech tagging, and
- Sentence boundary detection
Large-scale Structures: Exploiting Existing Features

- Explicit & Reoccurring Text Titles
- Chapter-boundaries
- Topical Outlines (sa bcad)
Case Example:

‘Gro-lung-pa’s Bstan-rim-chen-mo.

Full title from Title Page:

bdz bar gshegs pa'i bstan pa rin po che la 
'jug pa'i lam gyi rim pa rnam par bshad pa 
bzhugs so
Chapter Boundary Detection

Case Example:
‘Gro-lung-pa’s Bstan-rim-chen-mo.

Full title from Title Page:

bde bangshegs pa’i bstan pa rin po che la
jug pa’i lam gyi rim pa rnam par bshad pa
bzhugs so
Chapter Boundary Detection

Combine Title “Key” Syllables:

bde | gshegs | bstan | rin | po | che |
jug | lam | rim | rnam | bshad
Chapter Boundary Detection

With Chapter Colophon “Flags”:

de | ste | te | le’u | las
Chapter Boundary Detection

... and **Ordinal Numbers**:

```
(((nyer|nyi shu| ((sum|bzhi|lnga|drug|bdun|brgyad|dgu|brgya) (bcu|cu )?))?
  (((rtsa|so|zhe|nga|re|don|gya|go) )?))?)
((dang po| (((gcig|gnyis|gsum|bzhi|lnga|drug|bdun|brgyad|dgu|bcu|tham) ) + (pa)?))
```
Chapter Boundary Detection

Yields Automatic Colophon Identification:

TITLE + FLAG + ORDINAL

bstan pa la 'jug pa'i rim pa rnam pa bshad pa las dge ba'i bshes gnyen bsten pa la 'jug pa ste le'u dang po'o
Automatic Tagging of Large-scale Structures

...
Previous research explored three approaches:

- Statistical Methods
- Conceptual Hierarchies
- Exploiting lexical resources
Kozima (1993) put forth a method for calculating the Lexical Cohesion Profile (LCP) of English-language texts by:

- Building a weighted co-occurrence database of words from the *Longman Dictionary of Contemporary English*
- Performing a co-occurrence analysis over the text using a sliding Hanning window
LCP Method for Tibetan

- No resource comparable to Longman Dict. (Tshig-mdzod-chen-mo too uneven)
- Have two highly specialized genres of lit.:
  - Chos-kyi-rnam-grangs ("Enumerations of Phenomena")
  - Yig-cha ("Monastic Textbooks")
Chos-kyi-rnam-grangs

Sample Entry:

-'dus byas kyi mtshan nyid bzhi:
skye ba'i mtshan nyid
rga ba'i mtshan nyid
gnas pa'i mtshan nyid
mi rtag pa'i mtshan nyid do
Chos-kyi-rnam-grangs (stemmed & segmented)

Sample Entry:

\[ \text{\textquotesingle} \text{dus_byas kyi mtshan_nyid bzhi:} \]

\begin{itemize}
\item skye_ba mtshan_nyid
\item rga_pa mtshan_nyid
\item gnas_pa mtshan_nyid
\item mi_rtag_pa mtshan_nyid do
\end{itemize}
Yig-cha

Sample Entry:

☐ yid dpyod:

rang gi 'jug yul gyi gtso bor
gyur pa'i chos la 'tha' gcig tu zhen
kyang bcad don ma thob pa'i rig pa
Yig-cha
(stemmed & segmented)

Sample Entry:

☐ yid_dpyod:

rang_gi_jug_yul_gyi_gtso_bo
gyur_pa_chos_la_tha_gcig_tu_zhen
kyang_bcad_don_ma_thob_pa_rig_pa
Calculate TFIDF

- Term Frequency (TF) per entry, times Inverse Document Frequency (IDF) over the entire lexicon:

\[ \text{tfidf} = \log(1 + \log(tf)) \times \log(N/df) \]
Weighted & Normalized TFIDF

For example:

yid_dpyod (0.222390700)
rang_gi_jug_yul (0.166793025)
‘gyur_pa (0.008339651)
chos (0.011119535)
‘tha’_gcig (0.166793025)
etc ...
Hanning Weights

Rectangular Window

For window width, \( N \)
for \( 0 < n \leq N \), \( w(n) = 1 \)
else, \( w(n) = 0 \)

Hanning Window

For window width, \( N \)
\( w(n) = 1 - \cos \left( \frac{2\pi n}{N-1} \right) \)
Evaluation Metric: Known Item Retrieval

- Identify e-texts that have varied and rich vocabulary with known topic boundaries. Two test candidates — one canonical, one non-canonical:

  - Śāntideva’s Bodhicaryāvatāra
    (10 chapters; 26,887 syll.; 18,129 words)

  - Tsong-kha-pa’s Legs-bshad-snying-po
    (no chap. boundaries; 69,176 syll.; 42,956 words)
LCP for Śāntideva: Chos-kyi-rnam-grangs
LCP for Śāntideva: Yig-cha definitions
LCP for Prajñākaramati: Yig-cha definitions
LCP for Tsong-kha-pa: Chos-kyi-rnam-grangs
LCP for Tsong-kha-pa: Yig-cha definitions
LCP for Tsong-kha-pa: Yig-cha definitions
LCP for Tsong-kha-pa: Yig-cha definitions
Analysis

Immediate Observations:

1. Topic boundaries detection is feasible
2. Chapter boundaries are best / easily captured by non-CL methods
3. Chos-kyi-rnam-grangs fail, likely due to being “un-natural” lists
Applications

1. Fine grain indexing of texts based on individual sub-topics
2. Topic identification can be deployed for translation equivalent disambiguation
3. Content analysis and automatic topic outline generation easily done
Future Work

- Expand lexical Cohesion database with additional / alternate definitions
- Add domain tags to lexical pairs
- Incorporate domain tags in XML tagged documents for gisting/translation
fin.