Proceedings of the 12th Seminar of the International Association for Tibetan Studies,
Vancouver, BC, August 15-21, 2010 — Information Technology Panel

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

Paul G. Hackett
Columbia University
New York, New York, U.S.A.
ph2046@columbia.edu

Abstract

To properly implement a simple Tibetan Information Retrieval (IR) system segmentation of one form or another (n-gram, POS-tagging, dictionary substring matching, etc.) must be performed (see Hackett (2000b)). To take Tibetan indexing to a more sophisticated level however, some form of topic detection must be employed. This paper reports the results of a pilot study on the application to Tibetan of one technique for topic boundary detection: Lexical Cohesion. The resources developed and deployed, the theoretical model used, and its potential applications are discussed.

Introduction

In a previous paper (Hackett, 2000b) we demonstrated a method for performing word-segmentation in conjunction with part-of-speech tagging and sentence boundary detection. While sufficient for simple indexing and IR purposes, the assessment of larger scale structures within a text allows for more precise searching, translation equivalent disambiguation based on domain identification, and additional tagging possibilities. This paper reports the result of research deploying a method used by Kozima (1993) — "lexical cohesion" — for topic boundary detection, modified for Tibetan. Given the lack of comparable lexical resources for less-commonly studied languages like Tibetan, we exploit certain features in classical Tibetan literature, namely the literary genres of monastic textbooks (*yig cha*) and lists of enumerated phenomena (*chos kyi rnam grangs*), to build a keyword correlation database for use in computing "Lexical Cohesion Profiles" (LCP) for Tibetan texts.

Background

Previous research in topic boundary detection has tended to follow one of three approaches: statistical methods, utilizing conceptual hierarchies, or exploiting lexical resources.

In an approach utilizing statistical methods, Damashek (1995) reported success in the categorization of texts, although his results indicated that subtle differences (sub-topics) did not respond well to a statistical approach. Similarly, McHale assessed the clustering of documents through relying on hierarchical knowledge derived from resources such as WordNet in comparison with "flat" lexical resources, such as a thesaurus. Although distance-based similarity measures can be constructed from a conceptual hierarchy, McHale noted that the thesaurus approach tended to capture "the popular similarity of isolated word pairs" better than did methods relying upon such hierarchies.

Pursuing research in lexical resource methods, some have attempted to exploit mutual information in documents (for example, Fernández-Amorós, 2004), although excessive noise and false correlations have proved to hinder such approaches, requiring restriction of selected pairs. Even with lexical restriction modifications however, such an approach remains only moderately successful (Kan, et al., 1998; Fernández-Amorós, et al., 2010).

Lexical Cohesion for Large-scale structure / Boundary Detection

The concept of lexical cohesion was proposed by Halliday and Hasan (1976) as an indicator of the structure of a text. The method outlined by Morris and Hirst (1991) and applied by Kozima (1993) was an attempt to exploit the concept of lexical cohesion in a computational environment.

The approach is similar in nature to mutual information though dispenses with corpus-based statistics in favor of a fixed lexical resource. In Kozima's algorithm as applied to English, lexical cohesion between words was calculated on a semantic network constructed systematically from a subset of a standard monolingual English language dictionary. TDIDF weights were then computed and normalized to compute analog spreading activation (Waltz and Pollack, 1985) over the semantic network. From the network a Lexical Cohesion Profile (LCP) can be computed, which serves as a quantitative indicator of the smallest domain in which text coherence can be defined. The LCP is produced through calculating the density of lexical cohesion of words within a sliding Hanning window across the entire text. When plotted against word position, the resulting maxima and minima can be taken as a graphic representation of topics and topic boundaries (respectively) within a text.

Non-CL Methods for Large-scale structure / Boundary Detection for Tibetan

Without deploying computational linguistics methods such as lexical cohesion, it is possible to exploit several of the features of classical Tibetan literature to begin the process of finer-scale analysis and tagging of texts. Self-identified chapter boundaries and "topical outlines" (*sa bcad*) are two features of classical Tibetan literature that immediately lend themselves to such exploitation.

The title of a Tibetan text is often provided either at the start of a text (for canonical texts or those emulating that style) and/or in the closing colophon of a text, and is often clearly marked as such. By intelligently processing such text titles, a simple chapter-boundary detection pattern matching algorithm can be constructed (see Fig. 1).

Such an approach can also be taken for the purpose of automatically identifying "topical outlines" (*sa bcad*) within a Tibetan text, although given the issue of anaphora in typical instances, such a task requires greater sophistication.

```
1. Extract TITLE from DOCUMENT
2. Stem TITLE to remove all purely syntactic syllables

        s/( (gy?is?|kyis?|[pb]a(r|s|’i)?|[srdt]u|[rln]a) )/ /
        s/(\x27o )/ /
        s/(c|[sz]h)es bya/ /
        s/bzhugs (so)?/ /

3. Identify expected (FLAG) phrases:

        ((d|s?t)e |le\x27u |las )

and ordinal numbers (ORDINAL):

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

4. Construct a RegEx to capture variations on stemmed TITLE, FLAG, and ORDINAL

        TITLE + FLAG + ORDINAL + (rdzogs so|\x27o)
```

Fig. 1. Pseudo-code for Chapter-boundary Detection

Application of the Lexical Cohesion Method to Tibetan

In order to exactly replicate Kozima’s implementation of a lexical cohesion method utilizing lexical resources, the primary desired source would be either a well-formed mono-lingual dictionary or a thesaurus. Given the lack of such comparable lexical resources for less-commonly studied languages like Tibetan, instead we exploit certain genres in classical Tibetan literature. Two highly specialized genres of literature in the classical Tibetan corpus are the literary genres of monastic textbooks (*yig cha*) which contain philosophical definitions of terms, and lists of enumerated phenomena (*chos kyi rnam grangs*). Using these resources, two semantic networks were constructed for use in computing “Lexical Cohesion Profiles” (LCP) for Tibetan texts.

For the “Enumerated Phenomena” (*chos kyi rnam grangs*) semantic network, the popular text in this genre by the eighteenth century author, Kon-chok-jik-may-wang-po (*dkon mchog ‘jigs med dbang po*, 1728-1791) entitled “A Festival for the Minds of the Knowledgeable, An Enumeration of Phenomena Derived from Many Treatises of Sūtra and Tantra” (*mdo rgyud bstan bcos du ma nas ‘byung ba’i chos kyi rnam grangs shes ldan yid kyi dga’ ston*). Following culling of redundant and nested lists, the resulting data set contained roughly 500 list entries. For the “Monastic Textbooks” (*yig cha*) semantic network, roughly 240 philosophical definitions — slightly less than half the number of entries as the “Enumerated Phenomena” resource — were extracted from a work-in-progress (Hackett, in preparation). The data was stemmed and segmented (Hackett, 2000a), TFIDF weights calculated and normalized for each entry with headwords double-weighted. From each of these data sets a separate semantic network was constructed.

Evaluating the System

In choosing test texts for "known-item" evaluation (Reynar, 1994), two texts were chosen: one canonical text with explicit chapter boundaries, and one non-canonical text with explicit topical outline (*sa bcad*) boundaries.

The text chosen for chapter boundary identification was Śāntideva's "Guide to the Bodhisattva Way of Life" in its Tibetan translation, consisting of 18,129 words (26,887 syllables) and ten explicitly demarcated chapters. Though representative of canonical literature, since this text is predominantly in verse with terse grammar and vocabulary, a second test was also performed with the canonical commentary on the same text, Prajñākaramati's "Difficult Points Commentary on [Śāntideva's] 'Guide to the Bodhisattva Way of Life'." This latter text consists of 126,888 words (207,377 syllables) in nine explicitly demarcated chapters.

The text chosen for topical outline identification was the non-canonical text by Tsong-kha-pa, "The Essence of Eloquence" (*legs bshad snying po*), a philosophically complex text with an explicit embedded topical outline (*sa bcad*). The text is comprised of 69,176 syllables segmented into 42,956 words.

A Lexical Cohesion Profile (LCP) was generated for each text first using the "Enumerated Phenomena" (*chos kyi rnam grangs*) semantic network and then again with the "Monastic Textbooks" (*yig cha*) semantic network. Because of the sparse coverage in vocabulary in both semantic networks, a smoothing algorithm was applied to the resulting LCPs to produce averages at approximately 2% intervals across each text (250 word averages for the Śāntideva text, and 1,000 word averages for the Tsong-kha-pa text). The positions of known boundaries were then calculated for each text (see Appendix I) and plotted against each LCP.

Analysis

Test case: Śāntideva's *Bodhicaryāvatāra*

Evaluating the "Enumerated Phenomena" (chos kyi rnam grangs) semantic network against the Śāntideva text produced a LCP that did not appear to yield any of the known chapter boundaries (Fig. 2), and appeared to produce effectively random data.

The resulting LCP for the Śāntideva text produced by the "Monastic Textbooks" (yig cha) semantic network offered a different result. The resulting LCP appeared to yield six out of the nine known chapter boundaries — chapters 4, 5, 6, 7, 9, and 10 (Fig. 3) — or only 66% accuracy at the task. Furthermore, given the sparseness of the resulting profile, it is possible that some instances were artifacts of the incommensurability of the semantic network and subject matter of the text or portions thereof. A possible instance of this is the starting and ending boundaries for chapter 4, which had a null profile for the entire length of the chapter. Given that the subject matter of that chapter ("conscientiousness," bag yod) is not philosophical in nature, a null profile is not entirely unexpected.

To rule out the possibility that the observed boundaries were not mere artifacts of the short length of the text, a LCP for its major commentary — Prajñākaramati's Bodhicaryāvatāra-pañjikā in nine chapters — was also produced (Fig. 4). The resulting LCP appeared to yield five out of the eight known chapter boundaries — chapters 2, 3, 6, 7, 8, and 9 — or a similarly low 63% accuracy at the task.

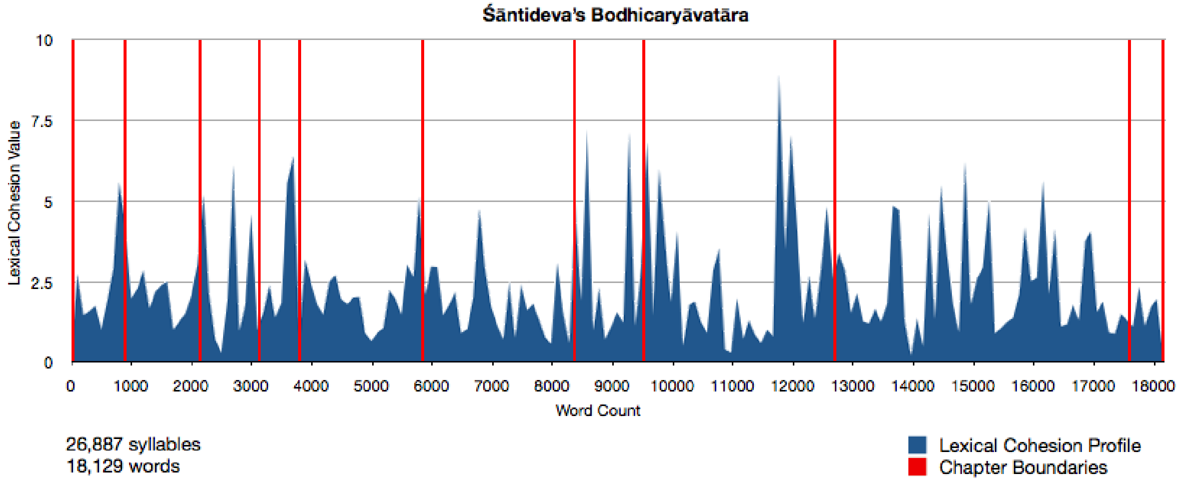

Fig. 2. LCP for Śāntideva's *Bodhicaryāvatāra* using Enumerated Phenomena Semantic Network

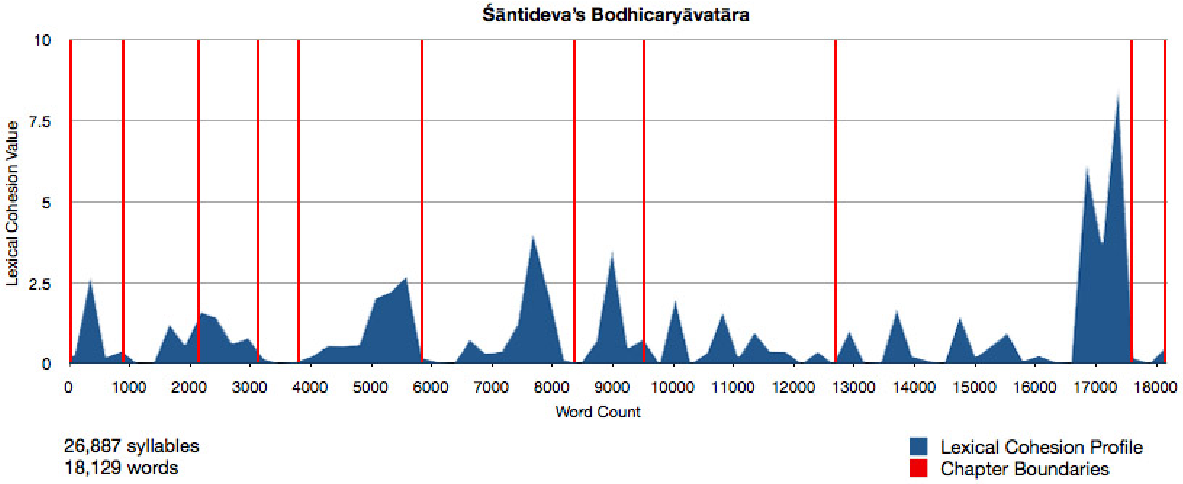

Fig. 3. LCP for Śāntideva's *Bodhicaryāvatāra* using Monastic Textbooks Semantic Network

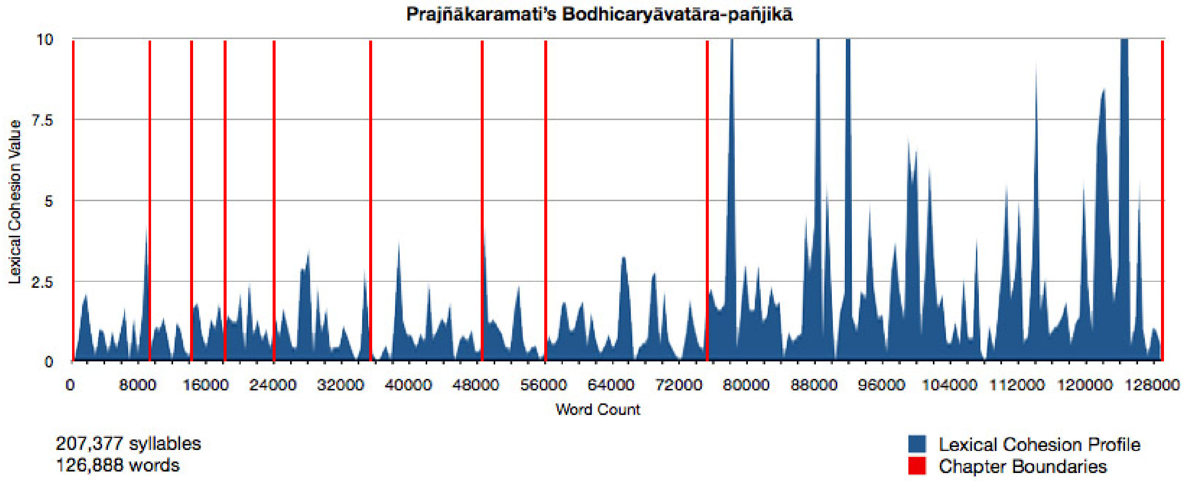

Fig. 4. LCP for Prajñākaramati's *Pañjika* using Monastic Textbooks Semantic Network

The conclusion drawn from these tests is that lexical cohesion appears to be only moderately successful at correctly identifying chapter boundaries, and that given the level of noise observed in the resulting LCPs, it would not be an efficient approach to that task.

Test case: Tsong-kha-pa's *Legs-bshad-snying-po*

Evaluating the "Enumerated Phenomena" (*chos kyi rnam grangs*) semantic network against the Tsong-kha-pa text produced similar results to the Śāntideva test, resulting in a LCP that did not appear to correspond to any of the major known topic boundaries (Fig. 5).

The resulting LCP for the Tsong-kha-pa text produced by the "Monastic Textbooks" (*yig cha*) semantic network likewise offered a different result. The two major divisions in the text were clearly observed (Fig. 6), while the six other minima in the LCP all corresponded to secondary divisions as well (Fig. 7). In addition, other local minima and shifts in the LCP appear to be indicative of subtle shifts in topic (Fig. 8), although should only be taken as suggestive.

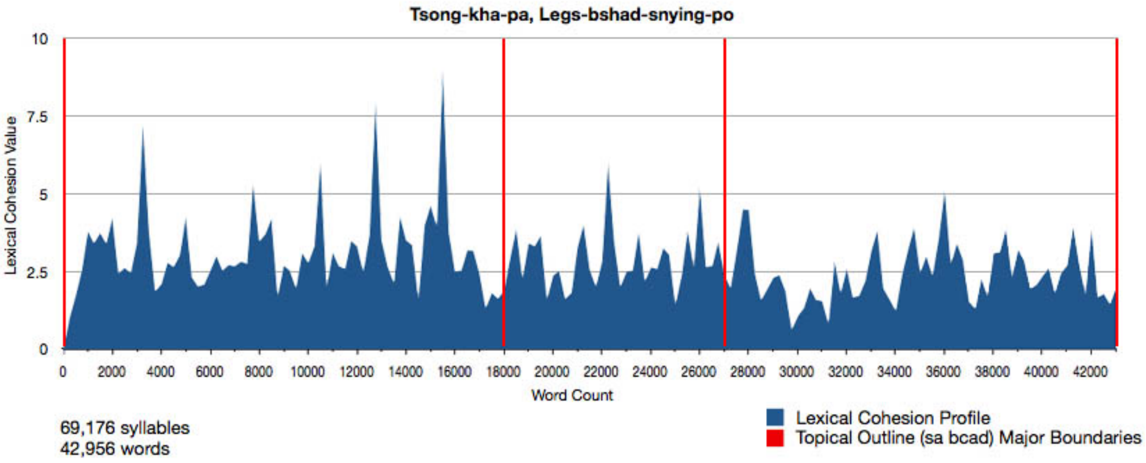

Fig. 5. LCP for Tsong-kha-pa's *Legs-bshad-snying-po* using Enumerated Phenomena Semantic Network

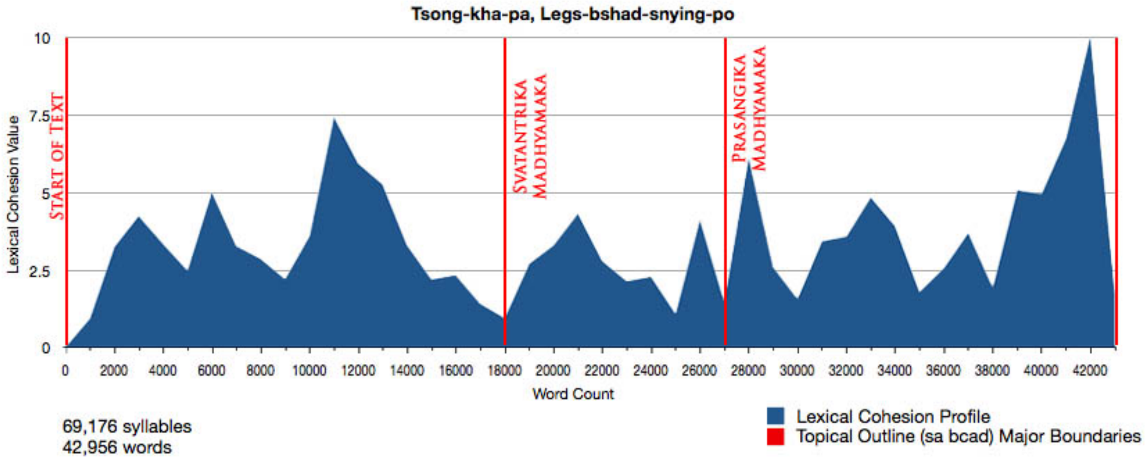

Fig. 6. LCP for Tsong-kha-pa's *Legs-bshad-snying-po* using Monastic Textbook Semantic Network

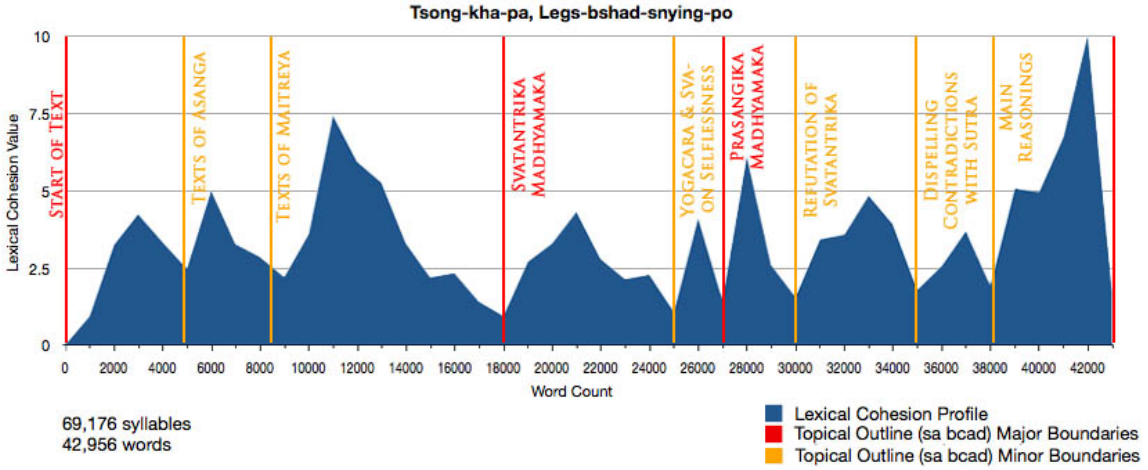

Fig. 7. LCP for Tsong-kha-pa's *Legs-bshad-snying-po* using Monastic Textbook Semantic Network

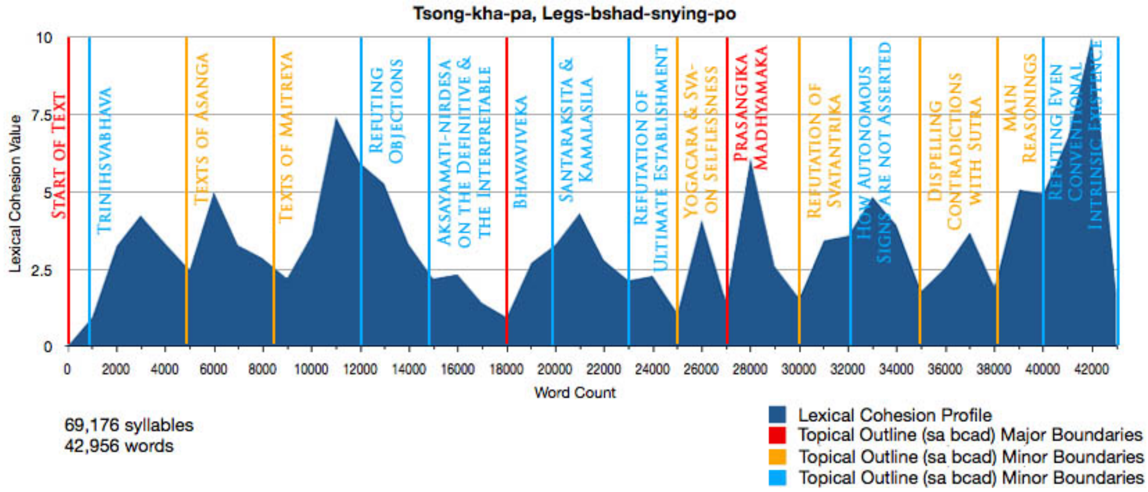

Fig. 8. LCP for Tsong-kha-pa's *Legs-bshad-snying-po* using Monastic Textbook Semantic Network

Conclusion

Summary

From an analysis of the performance of two different semantic networks against two distinctly different texts, three immediate observations that can be made:

1. Topic boundary detection is feasible for Tibetan with minimal lexical resources through the application of lexical cohesion methods,
2. While amenable to detection, chapter boundaries are best and easiest captured through the use of non-computation linguistic methods, such as pattern matching,
3. The use of resources from the genre of "Enumerated Phenomena" (*chos kyi rnam grangs*) for computing lexical cohesion does not appear to be warranted. The most likely reason for this failure would appear to be the "un-natural" nature of such lists, being compila-

tions extracted from individual texts or hypothetical constructions and not reflective of general principles. It remains possible that different sources together with lexical restriction could improve the accuracy of a semantic network construction on the basis of such works in this genre, but the manual overhead would likely be considerable.

Applications

The most immediate application of this technology is for the fine grain indexing of Tibetan texts based on individual sub-topics. As more and larger electronic text collections become available, the ability to perform more sophisticated searching and document retrieval becomes paramount. Related to this, is the need for cross-language information retrieval and gisting. With the ability to identify sub-topics in a text and the consequent assignment of domain labels, the ability to perform automatic translation equivalent disambiguation become feasible. With additional resources, gisting through lexical chaining (Yaari, 1997; Stokes, et al., 2004) and the automatic generation of topical outlines for otherwise undifferentiated texts can thus be performed.

Future Work

As a pilot study, we consider the basic premise of this research to have been proven to be sound. In order to broaden the applicability of this system to a larger range of texts and add functionality enabling the types of applications described above, we identify three courses of action to be taken:

- Expand the lexical cohesion database that serves as the basis for the semantic network by adding additional and alternate definitions from the textbook (*yig cha*) literature
- Add domain tags to the lexical pairs in order to enable sub-topic labeling
- Establish protocols for incorporating those resulting domain tags into XML-tagged documents for gisting and translation

References

Damashek, Marc. "Gauging Similarity with n-grams: Language-independent Categorization of Text," *Science* 276[10 Feb 1995]: 843-848.

Fernández-Amorós, David. "WSD Based on Mutual Information and Syntactic Patterns," in *Third International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL)*, 2004.

Fernández-Amorós, David, et al. "Automatic Word Sense Disambiguation Using Cooccurrence and Hierarchical Information," *Lecture Notes in Computer Science*, 6177 [2010]: 60-67.

Hackett, Paul G. "Approaches to Tibetan Information Retrieval:Segmentation vs. n-grams." Master's Thesis. College of Library and Information Services, University of Maryland, College Park (2000 (a)).

Hackett, Paul G. "Automatic Segmentation and Part-Of-Speech Tagging For Tibetan." Paper presented at the Ninth Seminar of the International Association for Tibetan Studies (IATS-9), Leiden, The Netherlands, June 2000 (b).

Hackett, Paul G. *Basic Buddhist Terms and Concepts*. Ithaca, NY: Snow Lion Publ. (in preparation).

Halliday, M.A.K., and R. Hasan. *Cohesion in English*. London: Longman (1976).

Kon-chok-jik-may-wang-po (*dkon mchog 'jigs med dbang po*, 1728-1791). "A Festival for the Minds of the Knowledgeable, An Enumeration of Phenomena Derived from Many Treatises of Sūtra and Tantra" (*mdo rgyud bstan bcos du ma nas 'byung ba'i chos kyi rnam grangs shes ldan yid kyi dga' ston*) in *Collected Writings of the Second Jamyang Zhepa Konchog Jigme Wangpo (1728-1791), Paramount Master of the Labrang Tashikyil Tradition*. New Delhi: Ngawang Gelek Demo, 1971; also, *chos kyi rnam grangs*. Kokonor, Tibet (China): Mtsho-sngon-mi-rigs-dpe-skrun-khang, 1989.

McHale, Michael. "A Comparison of WordNet and Roget's Taxonomy for Measuring Semantic Similarity" in *Proc. COLING/ACL Workshop Usage of WordNet in Natural Language Processing Systems*, 1998.

Kan, Min-Yen, Judith L. Klavans, and Kathleen R. McKeown. "Linear Segmentation and Segment Significance" in *Proceedings of the 6th International Workshop of Very Large Corpora (WVLC-6, Montreal)*, 1998, pp.197-205.

Morris, Jane and Graeme Hirst. "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," *Computational Linguistics* 17(1)[1991]: 21-48.

Prajñākaramati. "Difficult Points Commentary on [Śāntideva's] 'Guide to the Bodhisattva Way of Life'" (*bodhicaryāvatārapañjikā; byang chub kyi spyod pa la 'jug pa'i dka' 'grel*). Tōh. 3872. Derge Bstan-'gyur, Dbu-ma vol. LA, fol. 41b.1-288a.7.

Reynar, Jeffrey C. "An automatic method of finding topic boundaries" in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. June 27-30, 1994, Las Cruces, New Mexico, pp.331-333.

Śāntideva. "Guide to the Bodhisattva Way of Life" (*bodhicaryāvatāra; byang chub sems dpa'i spyod pa la 'jug pa*). Tōh. 3871. Derge Bstan-'gyur, Dbu-ma vol. LA, fol. 1b.1-40a.7.

Stokes, Nicola, et al. "Broadcast News Gisting Using Lexical Cohesion Analysis," *Lecture Notes in Computer Science*, 2997 [2004]: 209-222.

Waltz, David L., and Jordan B. Pollack. "Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation," *Cognitive Science* 9 [1985]: 51-74.

Yaari, Yaakov. "Segmentation of expository text by hierarchical agglomerative clustering" in *Recent Advances in NLP* (RANLP '97, Bulgaria).

Appendix I: Known Boundaries for Test Texts

I.a. Śāntideva's *Bodhicaryāvatāra*

Śāntideva's *Bodhicaryāvatāra* consists of ten chapter boundaries at the following word positions:

0	Start of Text / Chapter 1
750	Chapter 2
2080	Chapter 3
2739	Chapter 4
3734	Chapter 5
5820	Chapter 6
8482	Chapter 7
9999	Chapter 8
13660	Chapter 9
16902	Chapter 10
18047	End of Text

I.b. Prajñākaramati's *Bodhicaryāvatāra-pañjikā*

Prajñākaramati's *Bodhicaryāvatāra-pañjikā* consists of nine chapter boundaries at the following word positions:

0	Start of Text; Chapter 1
8865	Chapter 2
13964	Chapter 3
17661	Chapter 4
23420	Chapter 5
34722	Chapter 6
47643	Chapter 7
55213	Chapter 8
73965	Chapter 9
126825	End of Text

I.c. Tsong-kha-pa's *Legs-bshad-snying-po*

Tsong-kha-pa's *Legs-bshad-snying-po* presents a slightly different challenge since the embedded topical outline for the text is hierarchical, presenting different levels of granularity in topic boundaries. Consequently, the major divisions of the text were taken to be:

0	Start of Text; Cittamātra/Yogācara Section
17986	Svātantrika Mādhyamika Section
27170	Prāsaṅgika Mādhyamika Section
42956	End of Text

All other divisions in the text were taken to be secondary (asterisks indicate that the division was observed at a minimum in the LCP):

0	Prologue
864	Saṃdhinirmocana Sūtra and the Triniḥsvabhāva
* 4865	Texts of Asaṅga
7094	Dispelling that Contradictions Exist in Our Tenet System
* 8386	Texts of Maitreya
10128	Identifying Superimpositions
12283	Refuting Objections to Our Interpretation
13537	Differentiating Scriptures into the Definitive and the Interpretable
14787	Differentiating Scriptures based on the Akṣayamati-nirdeśa
*17986	Treatises of Bhavaviveka (Svātantrika Mādhyamika Section)
19846	Treatises of Śāntarakṣita and Kamalaśīla
23091	Śāntarakṣita and Kamalaśīla's Refutation of Ultimate Establishment
* 25551	Yogācara- & Svātantrika-Mādhyamika on Selflessness
* 27170	Coarse and Subtle Selflessness in the Śrāvaka Schools (Prāsaṅgika Mādhyamika Section)
29448	Not Asserting Self-cognizing Consciousnesses
* 30260	Refutation of Svātantrika
* 32072	How Autonomous Signs are not Asserted
34321	The Three Natures in the Madhyamaka School
* 34788	Dispelling Contradictions with Sūtra
* 38609	Main Prāsaṅgika Mādhyamika Reasonings
* 39114	Refuting even Conventional Intrinsic Existence
40938	Whether or not an Absence is proved by a Sign