Automatic Collective Commentaries with Corpus Positioning System.

ECAI Workshop 2017.3.6

Acedamia Sinica

yapcheahshen@gmail.com

Outline

Collective commentaries on paper

Layered Document and Corpus Positioning System.

Encoding External Markups with compact text-range.

Problem solved by layered Document.

Live Demostration

Collective commentaries on paper

Collect commentaries for a canonical text.

Break the commentaries by quote and locate it's origin.

Group the quotes by phrase of canonical text.

The relations between commentaries is a **mesh network**, Serialize them to paper is very tedious, costly and difficult to update.

欽定四庫全書 卷二千九百六十 經部

孟子說卷一 宋 張栻 撰

梁惠王上

孟子見梁惠王王曰叟不遠千里而來亦將有以利吾國乎孟子對曰王何必曰利亦有仁義而已矣王曰何以利吾國大夫曰何以利吾家士庶人曰何以利吾身上下交征利而國危矣萬乘之國弑其君者必千乘之家千乘之國弑其君者必百乘之家萬取千焉千取百

<-Origin

Commentaries->

孟子集註攷證卷一

宋金履祥撰 郡後學胡鳳丹月樵校

梁惠王上

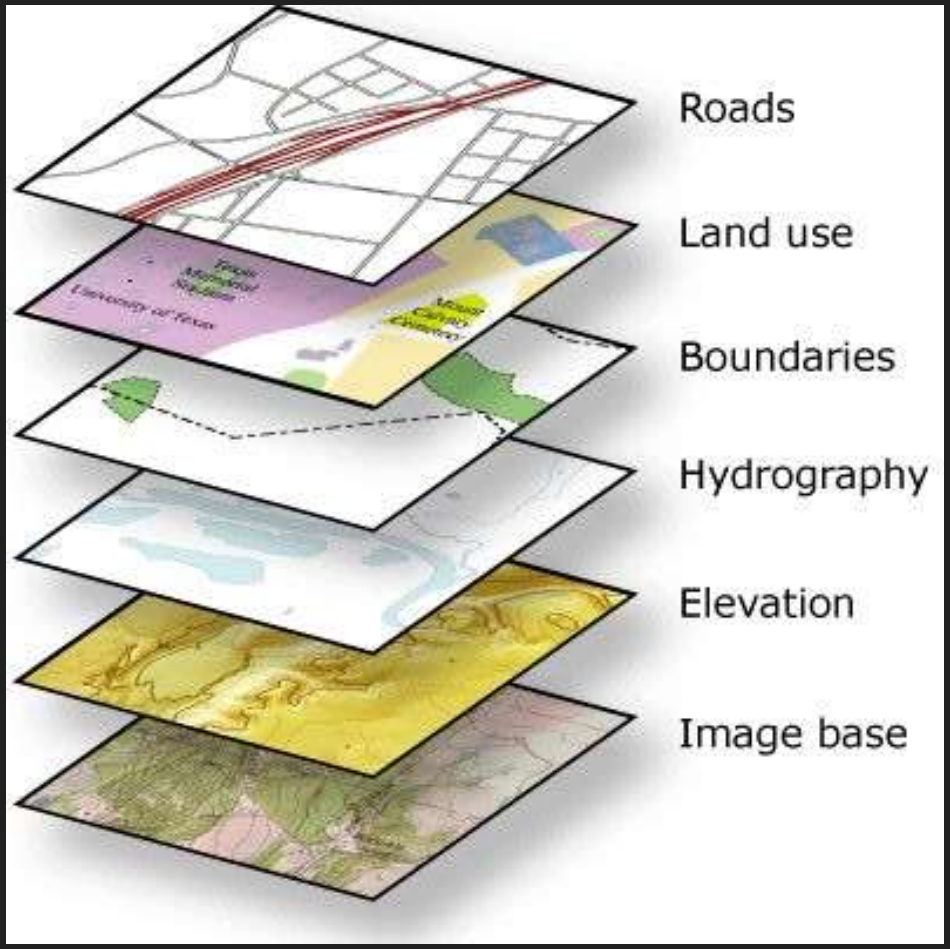
曉孟子每云不見諸侯而其書首云見梁惠王此固嫺倆之辭亦是不曾看史記史記云惠王數敗於軍旅卑禮厚幣以招賢者鄒衍淳于髡孟軻皆至梁那疏引此北士見之遂又謂孟子至梁鄒衍淳于髡從史記列傳稱鄒衍後孟子又云梁惠王郊迎鄒衍尊禮如此豈與孟軻困於齊梁同哉則梁惠之尊信孟子反不如衍此孟子道所以不行於梁也又傳雖稱客有見髡於梁惠王者然不云孟子見之也集註引史記是補孟子書之缺以知孟子之見梁惠王應其禮幣之聘爾叟宇當作叟俗作叟何文定謂當作叟俗作叟以叟老爲相尊之辭非必果有年也連一句又梁惠以叟稱孟子古人尚年一言率其下以利率下則上下交征國家必有簒奪之禍以仁義率下則下知仁義必無遺親後君之事而國家自無不利矣孟子此章分作兩節一節明言仁心之德是專言之仁義之利是偏言之明仁義之一節

Layered Document

XML store markups and text in same character string. This is good for ever-changing text (as markups are attached to text), but not suitable for collaboration and annotation.

Markups should be stored in separate layer on top of base text, base text is keep intact when adding or removing markups.

Thus we need a similar mechanism like GPS for corpus to connect markups and base text.



Corpus Positioning System (CPS)

Addressable down to word-level, every word/character can be identified precisely.

The simplest naive way is using the string-offset counting from 0, computer knows very well what is the the 1324151th character of Taisho Tripitaka
but this doesn't make any sense for human.

The address should be meaningful for human and machine.

Sparse encoding by extending legacy notation

Taisho legacy notation, **25p58a02** // volume 25, page 58 , column a , line 2

Volume, Page *Column , Line , Caret
100 , max 1100 page * 3 column, 29 lines, 19+1
|如|是|我|聞| // 4 characters, 5 carets

100<128 , 3300 <4096 , 29<32 , 20< 32
7 bits , 13 bits , 5 bits , 5bits = 30 bits

2^30=1,073,741,824(1 billion codes) maps to 150 millions concrete characters
Sparse: each characters has a unique code, but not vice versa.

Compact Text Range

A position in Taisho can be packed into 7,13, 5 ,5 = 30 bits

Assuming a text range will never cross a volume, thus it can be packed in 7+13+5+5 (start), 13+5+5(end, same volume) = 53 bits

53 bits = Maximum Integer percision of 64 bits floating point number

Markups=Text-Range+Arbitrary Payload (similar to tag attributes in XML)

Rich Document= Base text (with linebreak preserved) + Markups.

Benefits

As rich document is not a "tree" , the following problem are solved naturally:

- Not more xml:id and deeply nested tag.
- No more XML parsing process.
- Markups are "opt-in" and independent. (freedom of removing/adding tags)
- Relief from markups-more-than-text.

```
<text><body>
<cb:mulu level="1" type="分">2 分</cb:mulu><cb:mulu level="2" type="經">16 善生經</cb:mulu><cb:div type="fen">
<lb n="0072c08" ed="T"/>
<milestone n="12" unit="juan"/>
<lb n="0072c09" ed="T"/><cb:juan n="012" fun="open"><cb:mulu n="012" type="卷"></cb:mulu><cb:jhead><title><anchor xml:id="nkr_note_orig_0072016" n="0072016"/><anchor xml:id="beg0072016" n="0072016"/>佛說<anchor xml:id="end0072016"/>長阿含經</title>卷第十二</cb:jhead></cb:juan>
<lb n="0072c10" ed="T"/>
<lb n="0072c11" ed="T"/><byline cb:type="Translator"><anchor xml:id="nkr_note_orig_0072017" n="0072017"/><anchor xml:id="nkr_note_mod_0072017" n="0072017"/><anchor xml:id="beg0072017" n="0072017"/>後秦弘始年<anchor xml:id="end0072017"/>佛陀耶舍共竺佛念譯</byline>
<lb n="0072c12" ed="T"/><cb:div type="jing"><cb:mulu level="2" type="經">17 清淨經</cb:mulu><head>（一七）<anchor xml:id="nkr_note_orig_0072018" n="0072018"/><anchor xml:id="beg0072018" n="0072018"/>第<anchor xml:id="end0072018"/>二分<anchor xml:id="nkr_note_orig_0072019" n="0072019"/><anchor xml:id="nkr_note_equivalent_0072019" n="0072019"/>清淨經第十三</head>
<lb n="0072c13" ed="T"/><p xml:id="pT01p0072c1301">如是我聞：</p><p xml:id="pT01p0072c1305" rend="inline">一時，佛在迦維羅衛國緬祇優婆
<lb n="0072c14" ed="T"/>塞林中，與大比丘<anchor xml:id="beg_127" type="cb-app"/>眾<anchor xml:id="end_127"/>千二百五十人俱。</p><p xml:id="pT01p0072c1416" rend="inline">時，有沙
<lb n="0072c15" ed="T"/>彌<anchor xml:id="nkr_note_orig_0072020" n="0072020"/><anchor xml:id="nkr_note_mod_0072020" n="0072020"/><anchor xml:id="beg0072020" n="0072020"/>周那<anchor xml:id="end0072020"/>在<anchor xml:id="nkr_note_orig_0072021" n="0072021"/><anchor xml:id="nkr_note_mod_0072021" n="0072021"/><anchor xml:id="beg0072021" n="0072021"/>波波<anchor xml:id="end0072021"/>
國，夏安居已，執持衣鉢，
<lb n="0072c16" ed="T"/>漸詣迦維羅衛緬祇園中，至阿難所，頭面禮
<lb n="0072c17" ed="T"/>足，於一面立，白阿難言：「波波城內有<anchor xml:id="nkr_note_orig_0072022" n="0072022"/><anchor xml:id="nkr_note_mod_0072022" n="0072022"/><anchor xml:id="beg0072022" n="0072022"/>
```

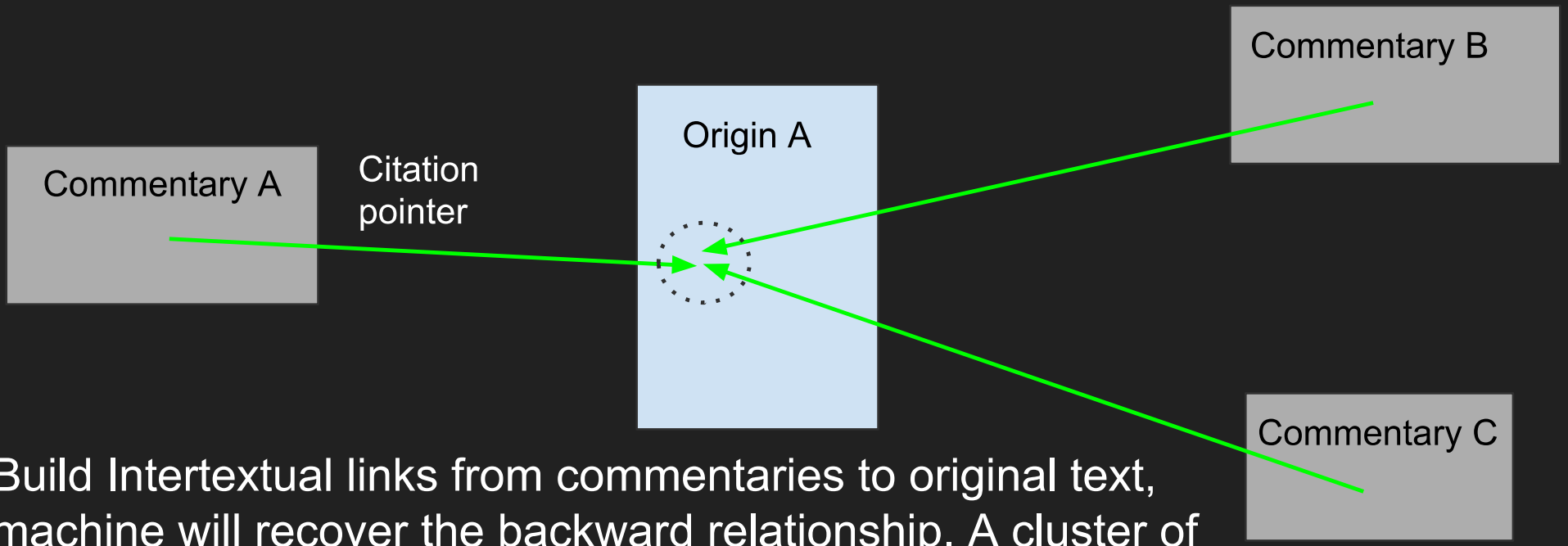
Beyond TEI/XML: Bi-directional Intertextual Links

Connecting sentence to any sentence.

HTML/XML can only link to beginning of the target document,
or predefined anchor in the target document
(require write access and maintain xml:id)

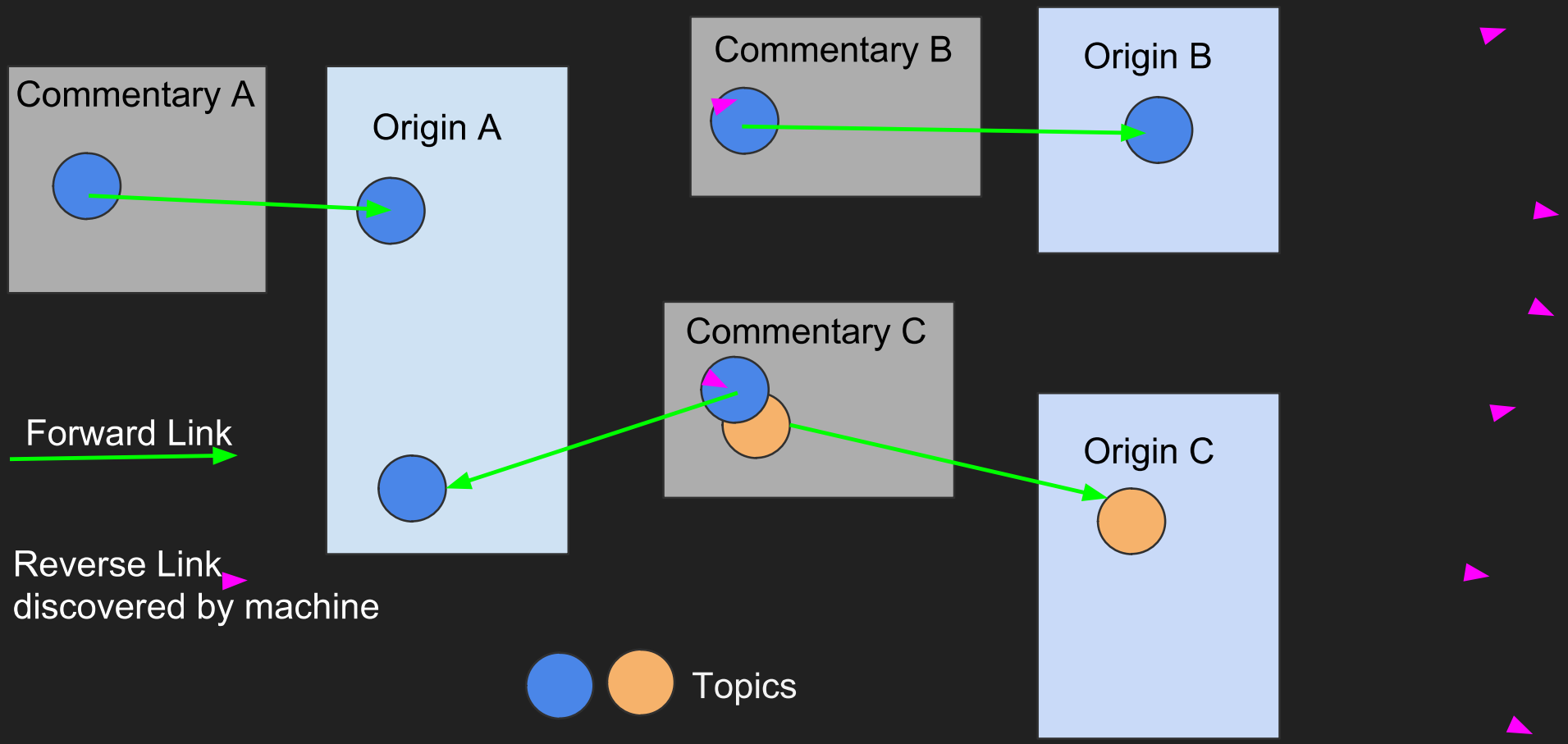
Same position might have many links, a new user interface is designed to follow multiple-target.

Automatic collective Commentaries



Build Intertextual links from commentaries to original text, machine will recover the backward relationship. A cluster of link is sketching the boundary of a topic.
Due to variants and punctuations, quote text and original are not identical, we use fuzzy search to locate the most possible text and ask human to confirm the link. DEMO 1

Chain Reaction of Topic Discovery



Live Demo

Building Bi-directional Intertextual Links
https://youtu.be/2JALMC6jT0E 1 minute

A new way of Topic Discovery
https://youtu.be/6W2d69YNylg 1.5 minutes

Thank you for your attention

yapcheahshen@gmail.com

This project is sponsered by Bodhiyana Foundation