Natural Language Processing Lab.
National Taiwan University

Lecture 13
Topic Tracking, Detection, and Summarization: Some IE Applications

Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
E-mail: hhchen@csie.ntu.edu.tw

Outline

- Topic Detection and Tracking
  - Topic Detection
  - Link Detection
- Summarization
  - Single Document
  - Multiple Document
  - Multilingual Document
- Summary

New Information Era

- How to extract the interesting information from large scale heterogeneous collection
- main technologies
  - natural language processing
  - information retrieval
  - information extraction

Topic Detection and Tracking (TDT)

Book:

Topic Detection and Tracking: Event-Based Information Organization, James Allan, Jaime Carbonnell, Jonathan Yamron (Editors), Kluwer, 2002

The TDT Project

History of the TDT Project

- Sponsor: DARPA
- Corpus: LDC
- Evaluation: NIST
- TDT Pilot Study -- 1997
- TDT phase 2 (TDT2) -- 1998
- TDT phase 3 (TDT3) – 1999
- …

TDT Tasks

- The Story Segmentation Task
- The First-Story Detection Task
- The Topic Detection Task
- The Topic Tracking Task
- The Link Detection Task

Topic

A Topic:

A topic is defined to be a seminal event or activity, along with all directly related events and activities.

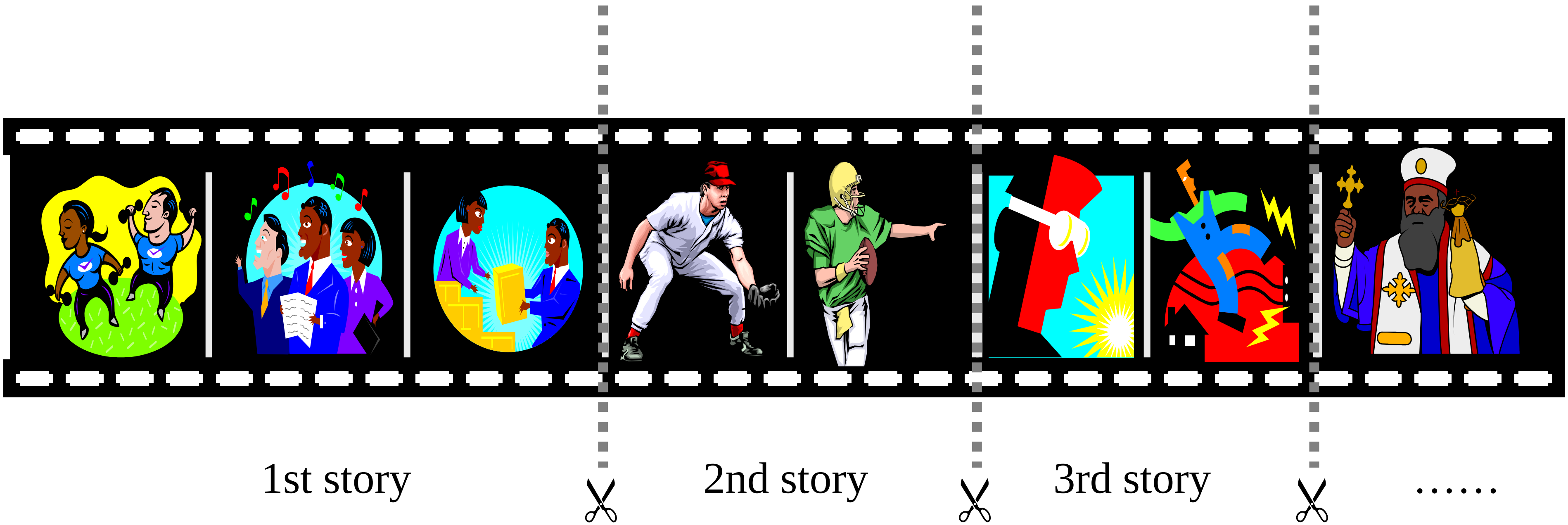
TDT3 topic detection task is defined as:

The task of detecting and tracking topics not previously known to the system

Topic Detection and Tracking (TDT)

- *Story Segmentation*
  - dividing the transcript of a news show into individual stories
- *First Story Detection*
  - recognizing the onset of a new topic in the stream of news stories
- *Cluster Detection*
  - grouping all stories as they arrive, based on the topics they discuss
- *Tracking*
  - monitoring the stream of news stories to find additional stories on a topic that was identified using several sample stories
- *Story Link Detection*
  - deciding whether two randomly selected stories discuss the same news topic

Story Segmentation



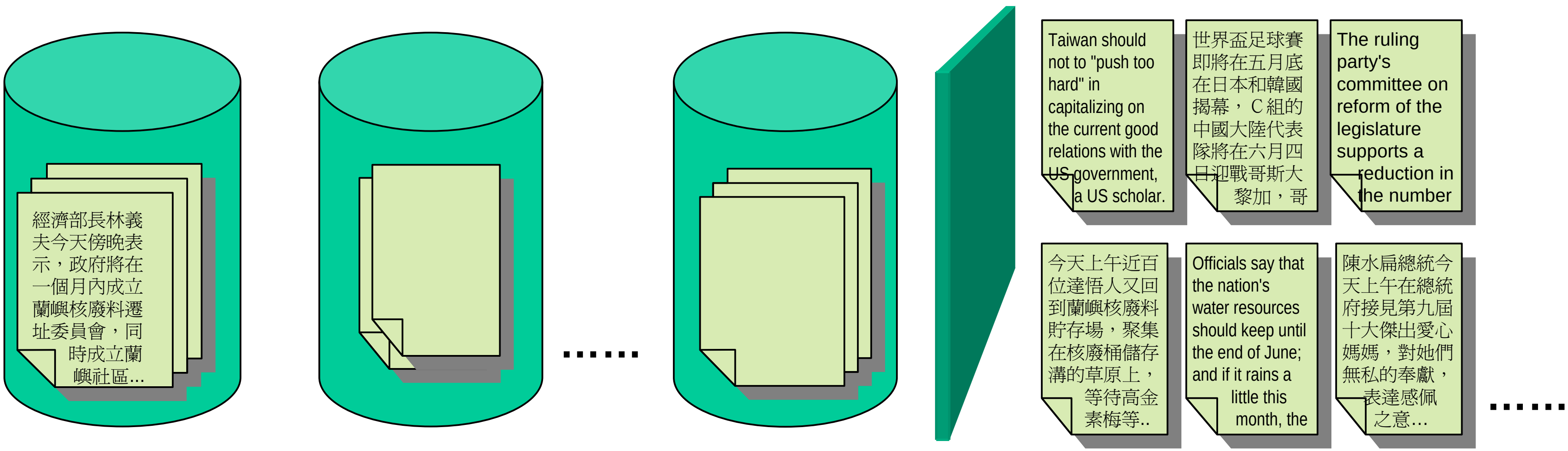
Story Segmentation

- goal
  - take a show of news and to detect the boundaries between stories automatically
- types
  - done on the audio source directly
  - using a text transcript of the show—either closed captions or speech recognizer output
- approaches
  - look for changes in the vocabulary that is used
  - look for words, phrases, pauses, or other features that occur near story boundaries, to see if they can find sets of features that reliably distinguish the middle of a story from its beginning or end, and clustering those segments to find larger story-like units

First Story Detection

- goal
  - recognize when a news topic appears that had not been discussed earlier
  - Detect that first news story that reports a bomb’s explosion, a volcano’s eruption, or a brewing political scandal
- approach
  - (1) Reduce stories to a set of features, either as a vector or a probability distribution.
  - (2) When a new story arrives, its feature set is compared to those of *all* past stories.
  - (3) If there is sufficient difference the story is marked as a first story; otherwise, not.
- applications
  - interest to information, security, or stock analysts whose job is look for new events that are of significance in their area

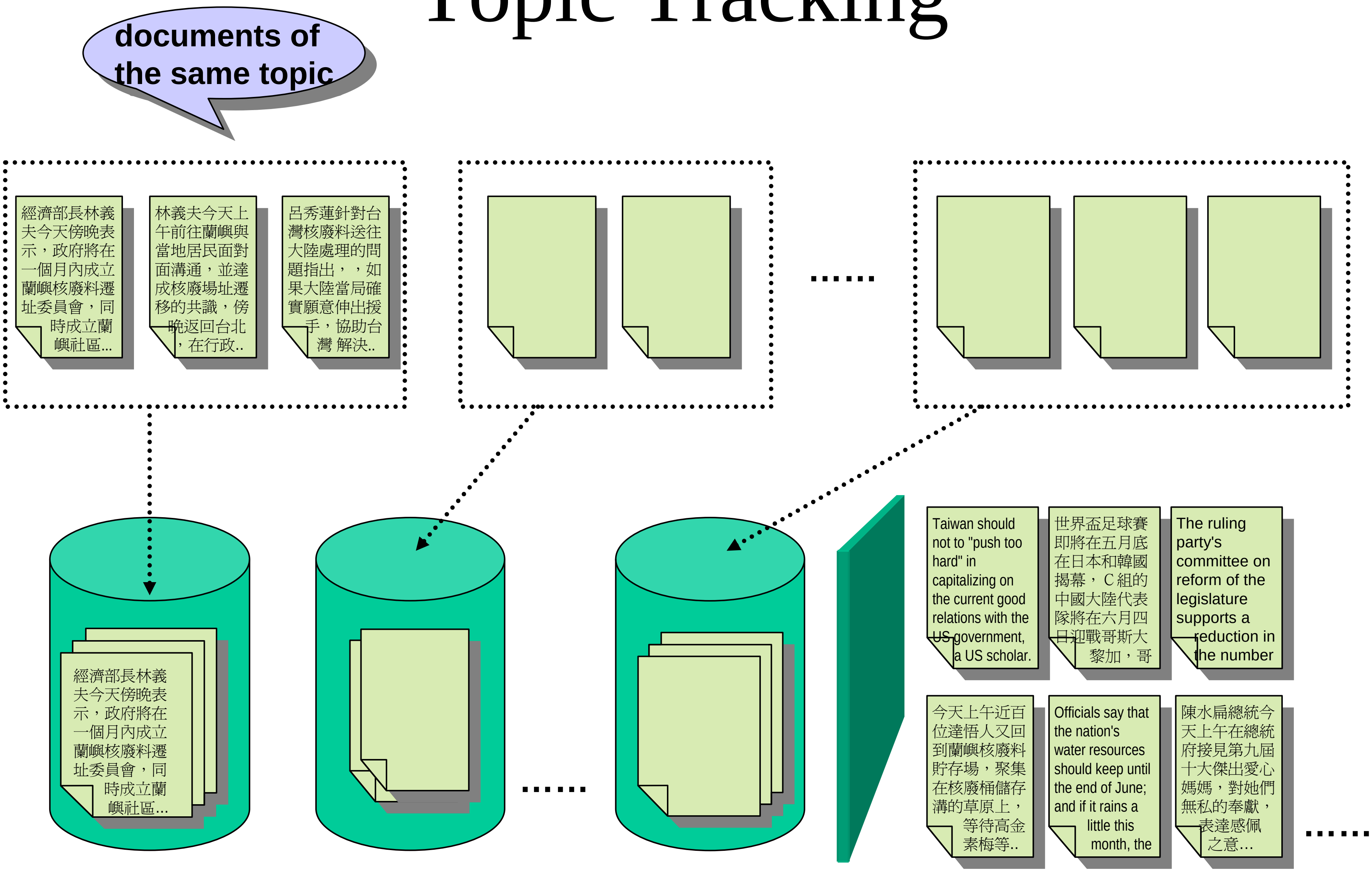
Cluster Detection



Cluster Detection

- goal
  - to cluster stories on the same topic into bins
  - the creation of bins is an unsupervised task
- approach
  - (1) Stories are represented by a set of features.
  - (2) When a new story arrives it is compared to all past stories and assigned to the cluster of the most similar story from the past (i.e., one nearest neighbor).

Topic Tracking

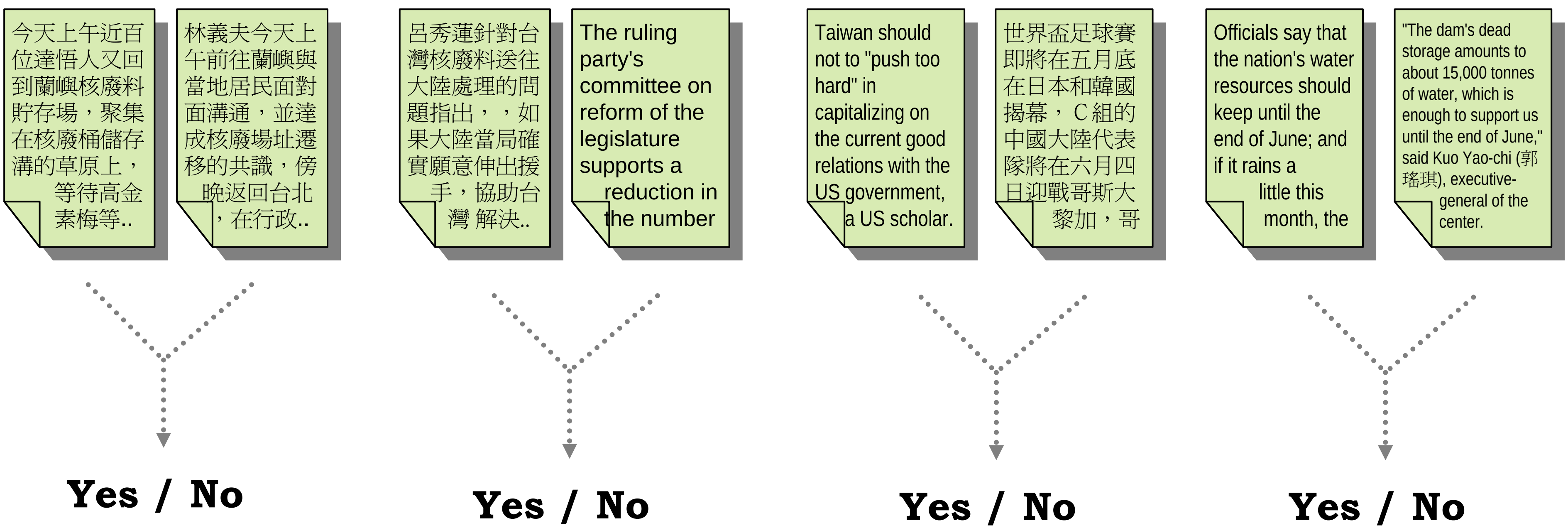


Tracking

- goal
  - similar information retrieval's filtering task
  - provided with a small number of stories that are known to be on the same topic, find all other stories on that topic in the stream of arriving news
- approach
  - extract a set of features from the training stories that differentiate it from the much larger set of stories in the past
  - When a new story arrives, it is compared to the topic features and if it matches sufficiently, declared to be on topic.

Story Link Detection

- goal
  - handed two news stories, determine whether or not they discuss the same topic



The TDT3 Corpus

- Source: Same as in TDT2 in English, VOA, Xinhua and Xaobao in Chinese.
- Total number of Stories: 34,600 (E), 30,000 (M)
- Total number of topics: 60 topics
- Time period: October - December, 1998
- Language type: English and Mandarin

Evaluation Criteria

- Use penalties
- Miss-False Alarm vs. Precision-Recall
- Cost Functions
- Story-weighted and Topic-weighted

Miss-False Alarm vs. Precision-Recall

	In topic	Not in topic
In topic (system)	(1)	(2)
Not in topic (system)	(3)	(4)

- Miss = (3) / [(1) + (3)]
- False alarm = (2) / [(2) + (4)]
- Recall = (1) / [(1) + (3)]
- Precision = (1) / [(1) + (2)]

Cost Functions

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{non-target}$$

$$(C_{Det})_{norm} = C_{Det} / MIN(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{non-target})$$

C_{Miss} (e.g., 10) and C_{FA} (e.g., 1) are the costs of a missed detection and a false alarm respectively, and are pre-specified for the application.

P_{Miss} and P_{FA} are the probabilities of a missed detection and a false alarm respectively and are determined by the evaluation results.

P_{Target} is the *a priori* probability of finding a target as specified by the application.

Cluster Detection

Hsin-Hsi Chen and Lun-Wei Ku (2002). “An NLP & IR Approach to Topic Detection.” *Topic Detection and Tracking: Event-Based Information Organization*, James Allan, Jaime Carbonnell, Jonathan Yamron (Editors), Kluwer, 243-264.

General System Framework

- Given a sequence of news stories, the topic detection task involves detecting and tracking topics not previously known to the system
- Algorithm
  - the first news story d_1 is assigned to topic t_1
  - assume there already are *k* topics when a new article d_i is considered
  - news story d_i may belong to one of *k* topics, or it may form a new topic t_{k+1}

How to make decisions

The first decision phase:

- Define similarity score S_{td}
- Relevant if $S_{td} > TH_{high}$
- Irrelevant if $S_{td} < TH_{low}$
- Undecided if $TH_{low} < S_{td} < TH_{high}$

The second decision phase:

- Define Medium threshold : $TH_{medium} = (TH_{high} + TH_{low}) / 2$
  Relevant if $S_{td} > TH_{medium}$
- Irrelevant if $S_{td} < TH_{medium}$

Deferral Period

- How long the system can delay when making a decision
- How many news articles the system can look ahead
- The “burst” nature of news articles
- The deferral period is defined in DEF
- DEF = 10

Issues

- (1) How can a news story and a topic be represented?
- (2) How can the similarity between a news story and a topic be calculated?
- (3) How can the two thresholds, i.e., TH_l and TH_h, be interpreted?
- (4) How can the system framework be extended to multilingual case?

Representation of News Stories

- Term Vectors for News Stories
  - the weight w_{ij} of a candidate term f_j in d_i

$$w_{ij} = \ln(tf_{ij}) \times idf_j$$

$$idf_j = \ln(\frac{n}{n_j})$$

  - tf_{ij} is the number of occurrences of f_j in d_i
  - n is the total number of topics that the system has detected
  - n_j is the number of topics in which f_j occurs
  - The first N (e.g., 50) terms are selected and form a vector for a news story

Representation of Topics

- Term Vectors for Topics
  - the time-variance issue: the event changes with time
  - d_i (an incoming news story) is about to be inserted into the cluster for t_k (the highest similarity with d_i)
  - Top-N-Weighted strategy
    - Select N terms with larger weights from the current V_{tk} and V_{di}
  - LRU+Weighting strategy
    - both recency and weight are incorporated
    - keep M candidate terms for each topic
    - N older candidate terms with lower weights are deleted
    - keep the more important terms and the latest terms in each topic cluster

Two Thresholds and the Topic Centroid

- The behavior of the centroid of a topic
- Define distance:

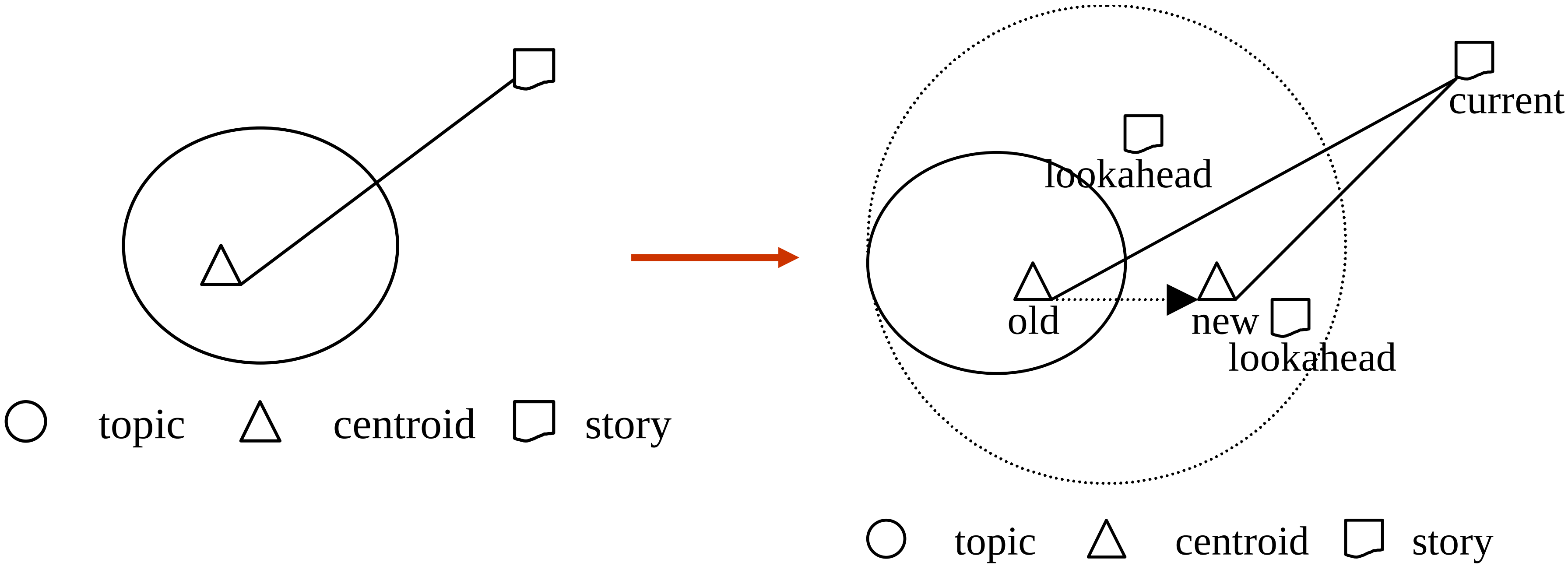
$$1 - S_{td}$$

$$S_{td} = \frac{<F_t> \cdot <F_d>}{|<F_t>| |<F_d>|}$$

- The more similar they are, the less the distance is.
- The contribution of relevant documents when look-ahead.

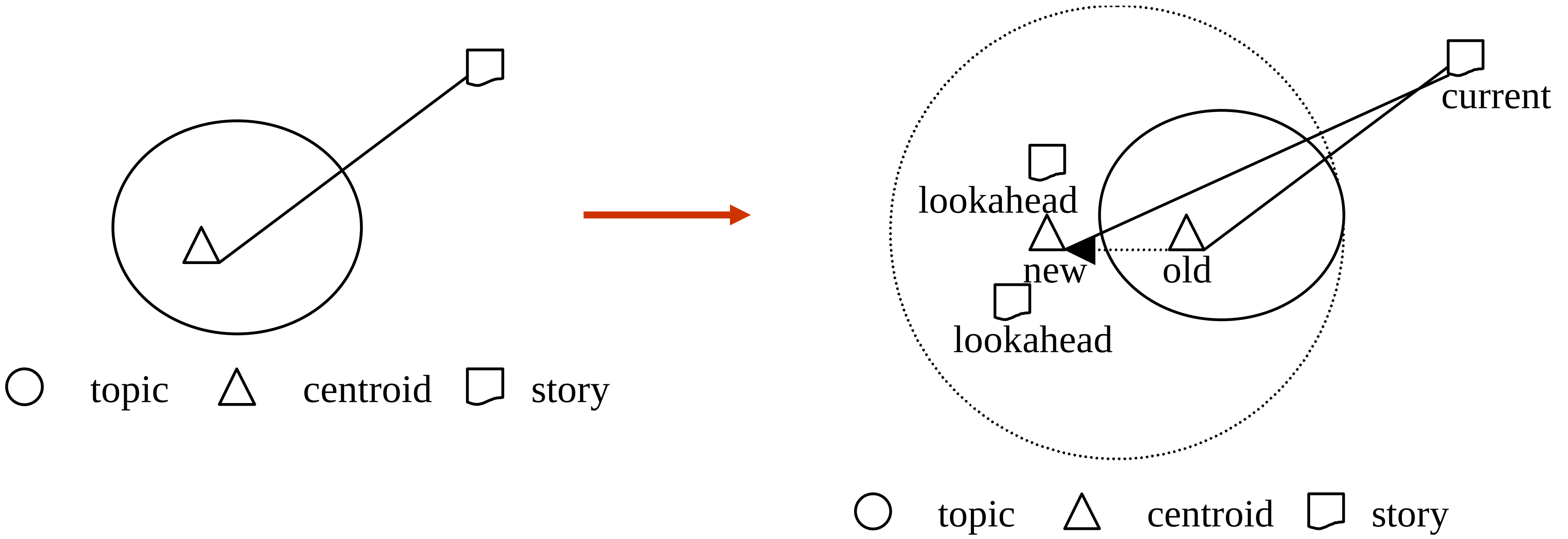
Two-Threshold Method

- relationship from undecidable to relevant



Two-Threshold Method

- Relationship from undecidable to irrelevant



Multilingual Topic Detection

- Lexical Translation
- Name Transliteration
- Representation of Multilingual News
  - For Mandarin news stories, a vector is composed of term pairs (Chinese-term, English-term)
  - For English news stories, a vector is composed of term pairs (nil, English-term)
- Representation of Topics
  - there is an English version (either translated or native) for each candidate term

Multilingual Topic Detection

- Similarity Measure
  - The incoming is a Mandarin news story
    - d_i is a represented as $<(c_{i1},e_{i1}), (c_{i2},e_{i2}), \ldots, (c_{iN},e_{iN})>$.
    - Use c_{ij} ($1 \le j \le N$) to match the Chinese terms in V_{tk}, and e_{ij} ($1 \le j \le N$) to match the English terms.
  - The incoming is an English news story
    - d_i is represented as $<(nil,e_{i1}), (nil,e_{i2}), \ldots, (nil,e_{iN})>$
    - Use e_{ij} ($1 \le j \le N$) to match the English terms in V_{tk}, and English translation of the Chinese terms.

Machine Transliteration

Classification

- Direction of Transliteration
  - Forward (Firenze➔翡冷翠)
  - Backward (阿諾史瓦辛格➔Arnold Schwarzenegger)
- Character Sets b/w Source and Target Languages
  - Same
  - Different

Forward Transliteration b/w Same Character Sets

- Especially b/w Roman Characters
- Usually no transliteration is performed.
- Example
  - Beethoven (貝多芬)
  - Firenze➔Florence, Muenchen➔Munich, Praha➔Prague, Moskva➔Moscow, Roma➔Rome
  - 小淵惠三

Forward Transliteration b/w Different Character Sets

- Procedure
  - Sounds in Source language➔Sounds in Target language➔Characters in Target language
- Example
  - 吳宗憲➔Wu × {Tsung, Dzung, Zong, Tzung} × {Hsien, Syan, Xian, Shian}
  - Lewinsky➔露文斯基, 柳思基, 陸雯絲姬, 陸文斯基, 呂茵斯基, 李文斯基, 露溫斯基, 蘿恩斯基, 李雯斯基, 李文絲基, *etc.*

Backward Transliteration b/w Same Character Sets

- Few or nothing to do because original transliteration is simple or straightforward

Backward Transliteration b/w Different Character Sets

- The Most Difficult and Critical
- Two Approaches
  - Reverse Engineering
  - Mate Matching

Similarity Measure

- In our study, transliteration is treated as similarity measure.
  - Forward: Maintain similarity in transliterating
  - Backward: Conduct similarity measurement with words in the candidate list

Three Levels of Similarity Measure

- Physical Sound
  - The most direct
- Phoneme
  - A finite set
- Grapheme

Grapheme-Based Approach

- Backward Transliteration from Chinese to English, a module in a CLIR system
- Procedure
  - Transliterated Word Sequence Recognition (i.e., named entity extraction)
  - Romanization
  - Compare romanized characters with a list of English candidates

Strategy 1: common characters

- How many common characters there are in a romanized Chinese proper name and an English proper name candidate.
- 埃斯其勒斯
- Wade-Giles romanization: ai.ssu.chi.le.ssu
- aeschylus
  ais suchilessu --> 3/9=0.33
- average ranks for a mate matching WG (40.06), Pinyin (31.05)

Strategy 2: syllables

- The matching is done in the syllables instead of the whole word.
- aes chy lus
  aissu chi lessu --> 6/9
- average ranks of the mate matching WG (35.65), Pinyin (27.32)

Strategy 3: integrate romanization systems

- different phones to denote the same sounds
  - consonants
    p vs. b, t vs. d, k vs. g, ch vs. j, ch vs. q,
    hs vs. x, ch vs. zh, j vs. r, ts vs. z, ts vs. c
  - vowels
    -ien vs. -ian, -ieh vs. -ie, -ou vs. -o,
    -o vs. -uo, -ung vs. -ong, -ueh vs. -ue,
    -uei vs. -ui, -iung vs. -iong, -i vs. -yi
- average ranks of mate matching: 25.39

Strategy 4: weights of match characters (1)

- Postulation:
  the first letter of each Romanized Chinese character is more important than others
- $score=\Sigma_i(f_i*(el_i/(2*cl_i)+0.5)+o_i*0.5)/el$
  el: length of English proper name,
  el_i: length of syllable i in English name,
  cl_i: number of Chinese characters corresponding to syllable i,
  f_i: number of matched first-letters in syllable i,
  o_i: number of matched other letters in syllable i

Strategy 4: weights of match characters (2)

- 埃斯 其 勒斯

埃斯	其	勒斯
aes	chy	lus
<u>A</u>i<u>S</u>su	<u>Ch</u>i	<u>L</u>e<u>S</u>su

- $el_1=3$, $cl_1=2$, $f_1=2$, $o_1=0$, $el=9$,
  $el_2=3$, $cl_2=1$, $f_2=1$, $o_2=1$,
  $el_3=3$, $cl_3=2$, $f_3=2$, $o_3=0$.
- average ranks of mate matching: 20.64
- penalty when the first letter of a Romanized Chinese character is not matched
  - average ranks: 16.78

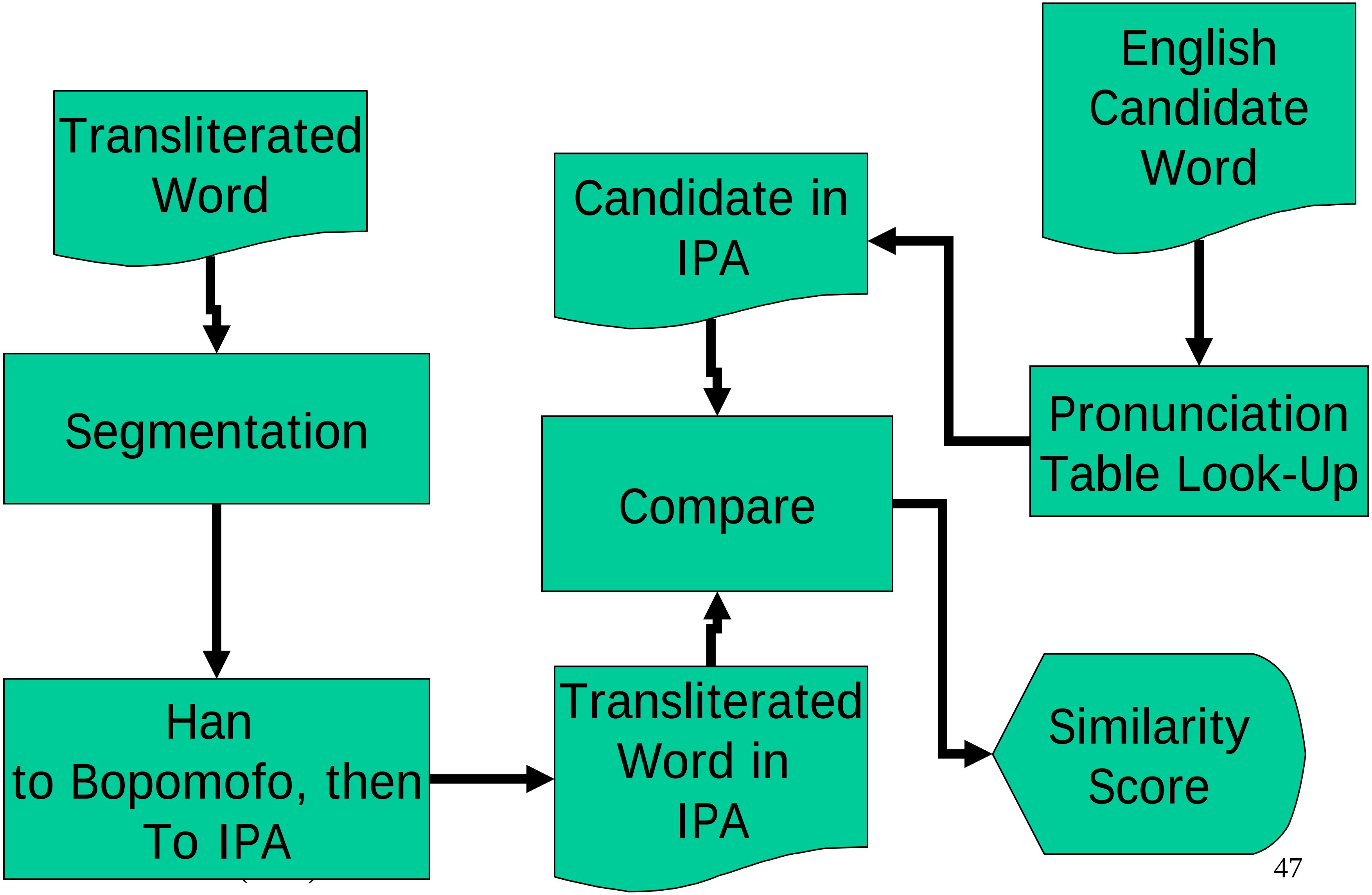
Strategy 5: pronunciation rules

- *ph* usually has *f* sound.
- average ranks of mate matching: 12.11
- performance of person name translation

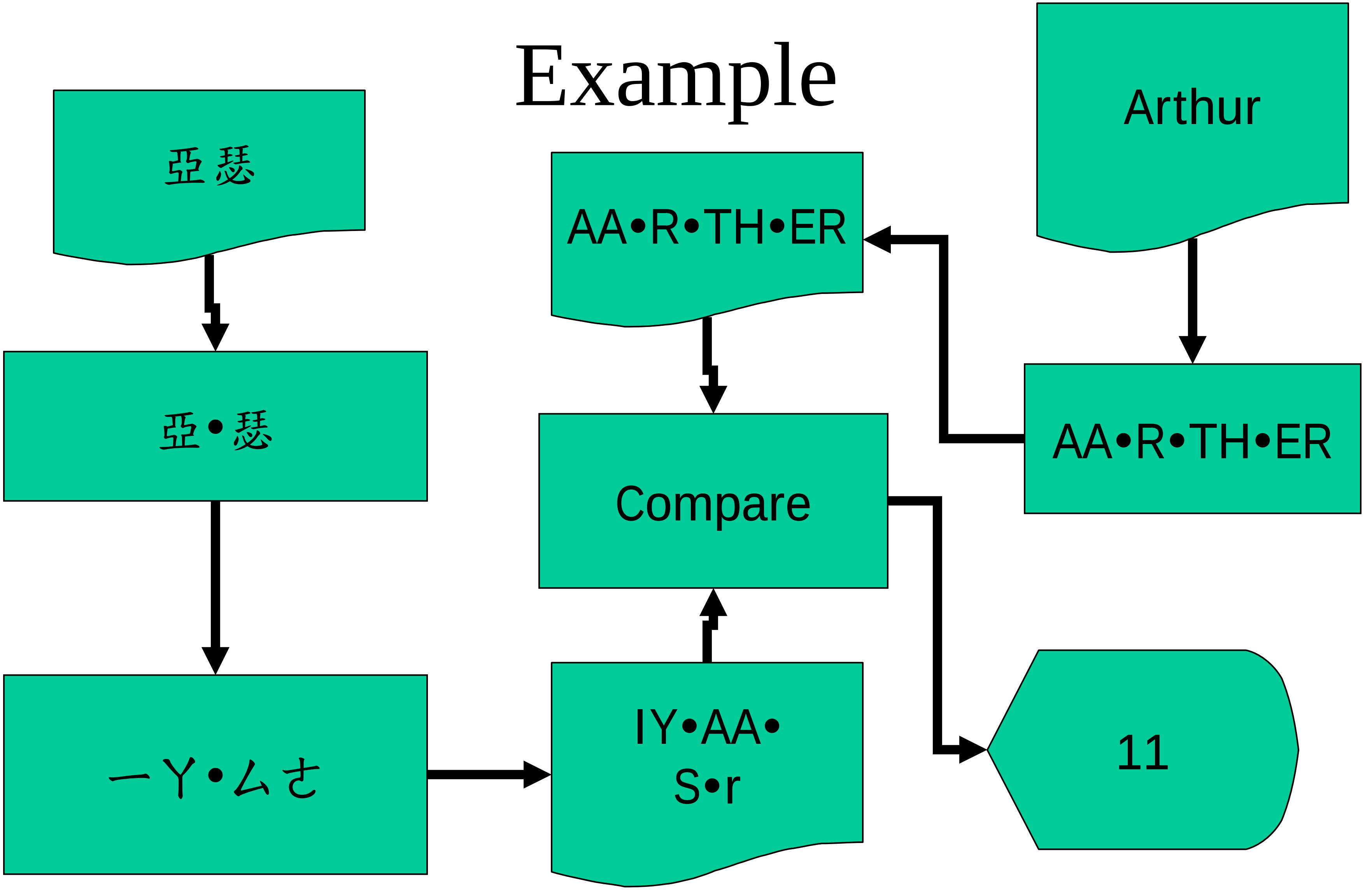
1	2-5	6-10	11-15	16-20	21-25	25+
524	497	107	143	44	22	197

- One-third have rank 1.

Phoneme-based Approach



Example



Similarity

- s(x, y) : similarity score between characters
- $\sum_{i=1}^{l} s(S_1^{'}(i), S_2^{'}(i))$: similarity score of an alignment of two strings
- Similarity score of two strings is defined as the score of the optimal alignment in given scoring matrix.

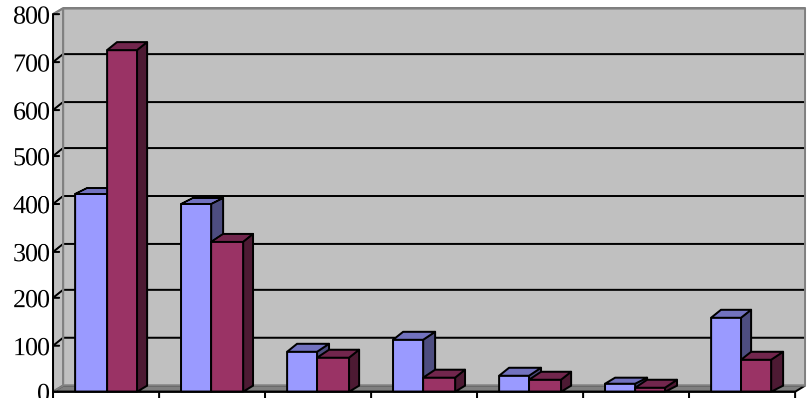
Compute Similarity

- Similarity can be calculated by dynamic programming in O(nm)
- Recurrence equation

$$
\begin{aligned}
V(i, j) = \max[&V(i-1, j-1) + s(S_1(i), S_2(j)), \\
&V(i-1, j) + s(S_1(i), \_), \\
&V(i, j-1) + s(\_, S_2(j))]
\end{aligned}
$$

Experiment Result

- Average Rank
  - 7.80 (Phoneme level) better than 9.69 (Grapheme level)
  - 57.65% is rank 1 (Phoneme level) > 33.28% (Grapheme level)



	1	2~5	6~10	11~15	16~20	21~25	25+
形素相似度	419	398	86	114	35	17	158
音素相似度	727	321	73	33	26	12	69

Experiments

Named Entities Only & the Top-N-Weighted Strategy

(Chinese Topic Detection)

TH_{low}	TH_{high}	Topic-Weighted P(miss)	Topic-Weighted P(F/A)	Cdet (norm)
0	0.20	0.6809	0.0075	0.7178
0.05	0.25	0.6884	0.0102	0.7385
0.10	0.30	0.6542	0.0068	0.6877
0.15	0.35	0.6717	0.0045	0.6938
0.20	0.40	0.6716	0.0037	0.6899

Named Entities Only & the LRU+Weighting Strategy (Chinese Topic Detection)

TH_{low}	TH_{high}	P(miss)	P(F/A)	Cdet (norm)	Change (vs. Table 3)
0.	0.20	0.6546	0.0014	0.6613	7.87%↑
0.05	0.25	0.6569	0.0008	0.6607	10.53%↑
0.10	0.30	0.6732	0.0003	0.6749	1.86%↑
0.15	0.35	0.6949	0.0002	0.6957	0.27%↓
0.20	0.40	0.7591	0.0001	0.7595	10.09%↓

The up arrow ↑ and the down arrow ↓ denote that the performance improved or worsened, respectively

Nouns-Verbs & the Top-N-Weighted Strategy (Chinese Topic Detection)

TH_{low}	TH_{high}	P(miss)	P(F/A)	Cdet (norm)	Change (vs. Table 3)
0	0.20	0.9739	0.0052	0.9993	39.21%↓
0.05	0.25	0.9946	0.0004	0.9965	34.94%↓
0.10	0.30	0.8745	0.0060	0.9039	31.44%↓
0.15	0.35	0.7943	0.0015	0.8015	15.52%↓
0.20	0.40	0.8119	0.0003	0.8134	17.90%↓

The performance was worse than that in the earlier experiments.

Nouns-Verbs & the LRU+Weighting Strategy
(Chinese Topic Detection)

TH_{low}	TH_{high}	P(miss)	P(F/A)	Cdet (norm)	Change (vs. Table 3)
0	0.20	0.5004	0.0025	0.5128	28.56%↑
0.05	0.25	0.5292	0.0015	0.5365	27.35%↑
0.10	0.30	0.6128	0.0008	0.6169	10.30%↑
0.15	0.35	0.6952	0.0003	0.6968	0.43%↓
0.20	0.40	0.7126	0.0002	0.7133	3.39%↓

The LRU+Weighting strategy was better than the top-N-weighted strategy when nouns and verbs were incorporated

Comparisons of Term and Strategies

	Named Entities Only	Nouns and Verbs
The top-N-weighted strategy	2	3
The LRU+Weighting strategy	2	1

Results with TDT-3 Corpus

TH_{low}	TH_{high}	Named Entities & LRU+W Cdet (norm)	Nouns-Verbs & LRU+W Cdet (norm)
0	0.20	0.5716	0.4327 (24.30%↑)
0.10	0.30	0.6166	0.4727 (23.34%↑)
0.15	0.35	0.6271	0.5610 (10.54%↑)
0.20	0.40	0.6812	0.4775 (29.90%↑)

English-Chinese Topic Detection

- A dictionary was used for lexical translation.
- For name transliteration, we measured the pronunciation similarity among English and Chinese proper names
  - A Chinese named entity extraction algorithm was applied to extract Chinese proper names
  - heuristic rules such as continuous capitalized words were used to select English proper names

Performance of English-Chinese Topic Detection

type	TH $_{low}$	TH $_{high}$	P(miss)	P(F/A)	Cdet (norm)
English-Chinese	0.1	0.2	0.5115	0.0034	0.5280
Chinese	0.1	0.3	0.4673	0.0011	0.4727

Named Entities

- Named entities, which denote people, places, time, events, and things, play an important role in a news story
- Solutions
  - Named Entities with Amplifying Weights before Selecting
  - Named Entities with Amplifying Weights after Selecting

Named Entities with Amplifying Weights before Selecting

amplification	TH $_{low}$	TH $_{high}$	P(miss)	P(F/A)	Cdet (norm)
weight $\times$ 1	0	0.15	0.4010	0.0060	0.4304
weight $\times$ 2	0	0.15	0.4335	0.0038	0.4519
weight $\times$ 3	0	0.15	0.4559	0.0032	0.4714

Named Entities with Amplifying Weights after Selecting

amplification	TH $_{low}$	TH $_{high}$	P(miss)	P(F/A)	Cdet (norm)
weight $\times$ 1	0	0.15	0.4010	0.0060	0.4304
weight $\times$ 2	0	0.15	0.3630	0.0027	0.3763
weight $\times$ 3	0	0.15	0.3552	0.0037	0.3740

Summarization

Information Explosion Age

- Large scale information is generated quickly, and crosses the geographic barrier to disseminate to different users.
- Two important issues
  - how to filter useless information
  - how to absorb and employ information effectively
- Example: an on-line news service
  - it takes much time to read all the news
  - personal news secretary
    - eliminate the redundant information
    - reorganize the news

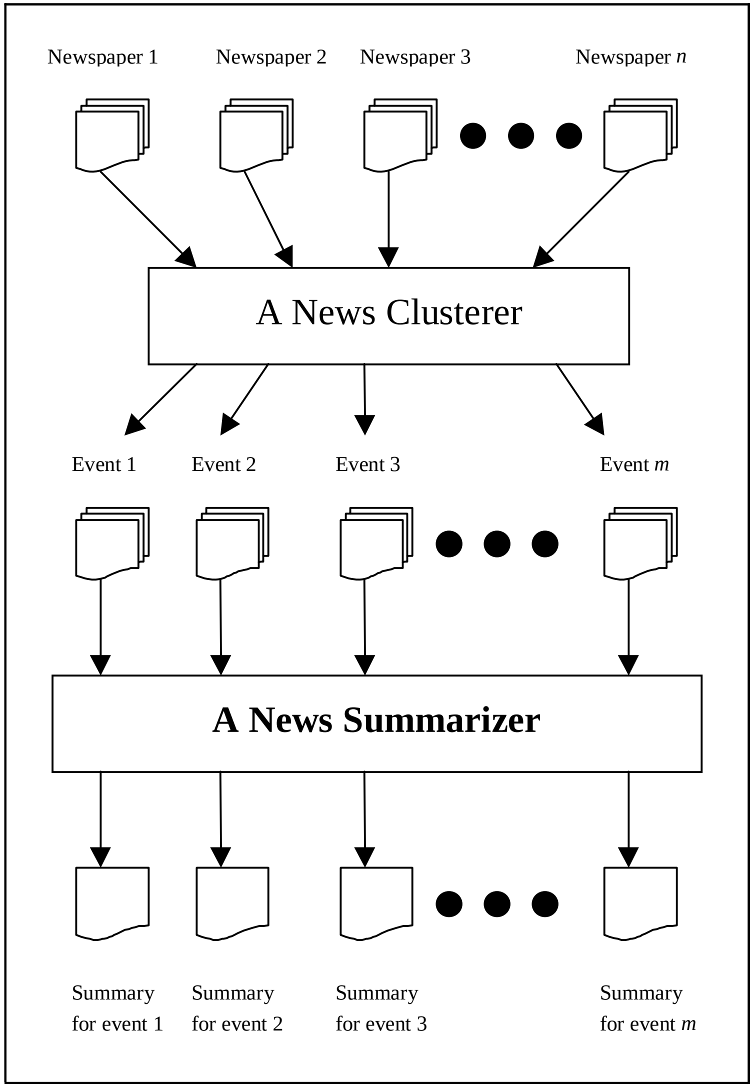
Summarization

- Create a shorter version for the original document
- applications
  - save users' reading time
  - eliminate the bottleneck on the Internet
  - …
- types
  - single document summarization
  - multiple document summarization
  - Multilingual multi-document summarization

Summac-1

- organized by DARPA Tipster Text Program in 1998
- evaluation of single document summarization
  - Categorization: Generic, indicative summary
  - Adhoc: Query-based, indicative summary
  - Q&A: Query-based, informative summary

Overview of our Summarization System



- Employing a segmentation system
- Extracting named entities
- Applying a tagger
- Clustering the news stream

- Partitioning a Chinese text
- Linking the meaningful units
- Displaying the summarization results

A News Clusterer (segmentation)

- identify the word boundary
- strategy
  - a dictionary
  - some morphological rules
    - numeral + classifier, e.g., 一個個，一條條
    - suffix, e.g., 學生們
    - special verbs, e.g.,吃吃 看,漂漂亮亮
  - an ambiguity resolution mechanism

A News Clusterer
(named entity extraction)

- extract named organizations, people, and locations, along with date/time expressions and monetary and percentage expressions
- strategy
  - character conditions
  - statistic information
  - titles
  - punctuation marks
  - organization and location keywords
  - speech-act and locative verbs
  - cache and *n*-gram model

Negative effects on summarization systems

- Two sentences denoting the similar meaning may be segmented differently due to the segmentation strategies.
  - 但法務部長城仲模內定升任司法院副院長 … --->
    但 法務部(Nc) 長城(Nc) 仲模(Nb) 內定(VC)升任(VG) 司法院(Nc) 副院長(Na) …
  - 而城仲模轉任司法院副院長之後的法務部長遺缺 --->
    而 城仲模(Nb) 轉任(VG) 司法院(Nc) 副院長(Na) 之後(Ng) 的 法務(Na) 部長(Na) 遺缺(Na)
  - major title and major person are segmented differently

Negative effects on summarization systems (*Continued*)

- Unknown words generate many single-character words
  - “土(Na) 石(Na) 流(VC)”, “園(Nc) 山(Na) 村(Nc)”, “芭(Nb) 比(VC) 絲(Na)”, “老(VH) 丙(Neu) 建(VC)”, and so on
- These words tend to be nouns and verbs, which are used in computing the scores for similarity measure.

A News Clusterer

- two-level approach
  - news articles are classified on the basis of a predefined topic set
  - the news articles in the same topic set are partitioned into several clusters according to named entities
- advantage
  - reducing the ambiguity introduced by famous persons and/or common names

Similarity Analysis

- basic idea in summarizing multiple news stories
  - which parts of new stories denote the same event?
  - what is a basic unit for semantic checking?
    - paragraph
    - sentence
    - others
- specific features of Chinese sentences
  - writers often assign punctuation marks at random
  - sentence boundary is not clear

Matching Unit

- example
  - 西班牙裔 是 美國 少數 族裔 人口 成長 最快 的 一 支 ， 這股 支持 力量 將 使 喬治 未來 在 與 共和黨 內 的 提名 競爭者 相較 之下 ， 別具 優勢 。
- matching unit
  - segments separated by comma
    - three segments
    - the segment may contain too little information
  - segments separated by period
    - one segment
    - the segment may contain too much information

Meaningful Units

- linguistic phenomena of Chinese sentences
  - about 75% of Chinese sentences are composed of more than two segments separated by commas
  - a segment may be an S, a NP, a VP, an AP, or a PP
- Meaningful unit is a basic matching unit
- previous example
  - 西班牙裔 是 美國 少數 族裔 人口 成長 最快 的 一 支
  - 這 股 支持 力量 將 使 喬治 未來 在 與 共和黨 內 的 提名 競爭者 相較 之下 ， 別 具 優勢

Meaningful Units (*Continued*)

- a MU that is composed of several sentence segments denotes a complete meaning
- three criteria
  - punctuation marks
    - sentence terminators: period, question mark, exclamation mark
    - segment separators: comma, semicolon and caesura mark

Meaningful Units (*Continued*)

- linking elements
  - forward-linking
    - a segment is linked with its next segment
    - 下課<u>之後</u>，我要去看電影。
(<u>After</u> I get out of class, I want to see a movie.)
  - backward-linking
    - a segment is linked with its previous segment
    - 我本來想去看電影，<u>可是</u>我沒有買到票。
(Originally, I had intended to see a movie, <u>but</u> I didn't buy a ticket.)
  - couple-linking
    - two segments are put together by a pair of words in these two segments
    - <u>因為</u>我沒有買到票，<u>所以</u>我沒有去看電影。
(<u>Because</u> I didn't by a ticket, (<u>so</u>) I didn't see a movie.)

Meaningful Units (*Continued*)

- topic chain
  - The topic of a clausal segment is deleted under the identity with a topic in its preceding segment
  - 他駕駛這艘太空梭，*e* 在太空中繞著月球飛行，*e* 等待這兩個人完成工作。
    (He drove the space shuttle and *e* flew around the moon, *e* waiting for these two men completing their jobs)
  - given two VP segments, or one S and one VP segments, if their expected subjects are unifiable, then the two segments can be linked (Chen, 1994)
  - We employ part of speech information only to predict if a subject of a verb is missing. If it does, it must appear in the previous segment and the two segments are connected to form a larger unit.

Similarity Models

- basic idea
  - find the similarity among MUs in the news articles reporting the same event
  - link the similar MUs together
  - verbs and nouns are important clues for similarity measures
  - example (nouns: 4/5, 4/4; verbs: 2/3, 2/2)

國父紀念館(Nc) 展出(VC) 的 國父(Na) 書畫(Na) 、 墨寶(Na)，傳出(VC) 了 失蹤(VH) 的 消息(Na)

國父紀念館(Nc) 傳出(VC) 了 國父(Na) 書畫(Na) 失蹤(VH) 的 消息(Na)

Similarity Models (*Continued*)

- strategies
  - (S1) Nouns in one MU are matched to nouns in another MU, so are verbs.
  - (S2) The operations in (S1) are exact matches.
  - (S3) A Chinese thesaurus is employed during the matching.
  - (S4) Each term specified in (S1) is matched only once.
  - (S5) The order of nouns and verbs in MU is not considered.
  - (S6) The order of nouns and verbs in MU is critical, but it is relaxed within a window.
  - (S7) When continuous terms are matched, an extra score is added.
  - (S8) When the object of transitive verbs are not matched, a score is subtracted.
  - (S9) When date/time expressions and monetary and percentage expressions are matched, an extra score is added.

Testing Corpus

- Nine events selected from Central Daily News, China Daily Newspaper, China Times Interactive, and FTV News Online
  - 社會役的實施 (military service): 6 articles
  - 老丙建建築 (construction permit): 4 articles
  - 三芝鄉土石流 (landslide in Shan Jr): 6 articles
  - 總統布希之子 (Bush's sons): 4 articles
  - 芭比絲颱風侵台 (Typhoon Babis): 3 articles
  - 股市穩定基金 (stabilization fund): 5 articles
  - 國父墨寶失竊案 (theft of Dr Sun Yat-sen's calligraphy): 3 articles
  - 央行調降利率 (interest rate of the Central Bank): 3 articles
  - 內閣總辭問題 (the resignation issue of the Cabinet): 4 articles

Experiment Results

- Model 1 (baseline model)
  - (S1) Nouns in one MU are matched to nouns in another MU, so are verbs.
  - (S3) The operations in (S1) is relaxed to inexact matches.
  - (S4) Each term specified in (S1) is matched only once.
  - (S5) The order of nouns and verbs in MU is not considered.
- Precision: 0.5000, Recall: 0.5434
- Consider the subject-verb-object sequence
  - The matching order of nouns and verbs are kept conditionally

Experiment Results (*Continued*)

- Model 2 = Model 1 - (S5) + (S6)
  - (S5) The order of nouns and verbs in MU is not considered.
  - (S6) The order of nouns and verbs in MU is critical, but it is relaxed within a window.
  - M1 precision: 0.5000 recall: 0.5434
  - M2 precision: 0.4871 ↓ recall: 0.3905 ↓
  - The syntax of Chinese sentences is not so restricted
- Give up the order criterion, but we add an extra score when continuous terms are matched, and subtract some score when the object of a transitive verb is not matched.

Experiment Results (*Continued*)

- Model 3 = Model 1 +
  - (S7) When continuous terms are matched, an extra score is added.
  - (S8) When the object of transitive verbs are not matched, a score is subtracted.
  - M1 precision: 0.5000 recall: 0.5434
  - M2 precision: 0.4871 ↓ recall: 0.3905 ↓
  - M3 precision: 0.5080 ↑ recall: 0.5888 ↑
- Consider some special named entities such as date/time expressions and monetary and percentage expressions

Experiment Results (*Continued*)

- Model 4 = Model 3 +
  - (S9) When date/time expressions and monetary and percentage expressions are matched, an extra score is added.
  - M1 precision: 0.5000 recall: 0.5434
  - M2 precision: 0.4871 ↓ recall: 0.3905 ↓
  - M3 precision: 0.5080 ↑ recall: 0.5888 ↑
  - M4 precision: 0.5164 ↑ recall: 0.6198 ↑
- Estimate the function of the Chinese thesaurus

Experiment Results (*Continued*)

- Model 5 = M4 - (S3) + (S2)
  - (S3) The operations in (S1) is relaxed to inexact matches.
  - (S2) The operations in (S1) are exact matches.
  - M4 precision: 0.5164 recall: 0.6198
  - M5 precision: 0.5243 ↑ recall: 0.5579 ↓

Analysis

- The same meaning may not always be expressed in terms of the same words or synonymous words.
- We can use different format to express monetary and percentage expressions.
  - two hundreds and eighty-three billions
    二千八百三十億元，二八三〇億元，2830億
  - seven point two five percent
    百分之七點二五，七•二五％” or ”7.25%
- segmentation errors
- incompleteness of thesaurus
  - Total 40% of nouns and 21% of verbs are not found in the thesaurus.

Presentation Models

- display the summarization results
  - browsing model
    - the news articles are listed by information decay
  - focusing model
    - a summarization is presented by voting from reporters

Browsing Model

- The first news article is shown to the user in its whole content.
- In the news articles shown latter, the MUs denoting the information mentioned before are shadowed.
- The amount of information in a news article is measured in terms of the number of MUs.
- For readability, a sentence is a display unit.

今日新聞總表 - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 下一頁 停止 重新整理 首頁 搜尋 我的最愛 記錄 郵件 字型 列印 編輯 討論

網址(D) http://nlg3.csie.ntu.edu.tw/systems/summary/main.html 移至 連結

今日新聞總表

2000年10月15日

[系統簡介]

■重點新聞

總統保證絕不加稅八大結論振興財經	8篇 [瀏覽式摘要] [重點式摘要]
張俊雄說明財經會議結論	4篇 [瀏覽式摘要] [重點式摘要]
僞卡解碼盜光存款	4篇 [瀏覽式摘要] [重點式摘要]

■政治新聞

大選期間中資援扁？國臺辦：荒唐說法無中生有	3篇 [瀏覽式摘要] [重點式摘要]
北高兩市議員除外地方民代支給大幅調高	3篇 [瀏覽式摘要] [重點式摘要]
國民黨祭出第三波黨紀懲處	2篇 [瀏覽式摘要] [重點式摘要]

■社會新聞

九組分工廣告攬客地下錢莊月放款逾億元	4篇 [瀏覽式摘要] [重點式摘要]
官田菱角節貴客臨門	4篇 [瀏覽式摘要] [重點式摘要]
東海餐會募得千餘萬	3篇 [瀏覽式摘要] [重點式摘要]

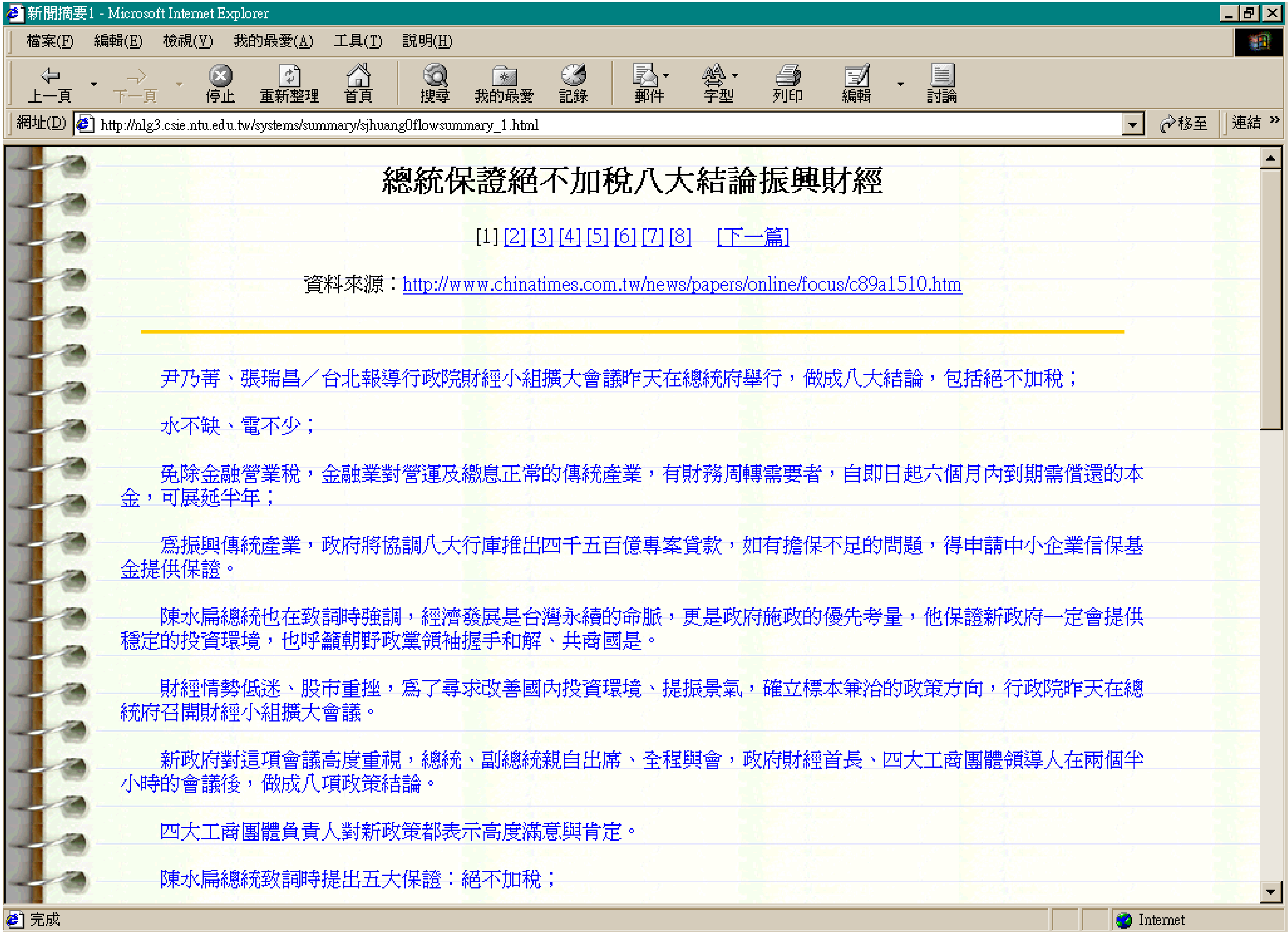
■地方新聞

憂鬱症篩檢日　高北同步解憂	3篇 [瀏覽式摘要] [重點式摘要]
環東大道西段通了	3篇 [瀏覽式摘要] [重點式摘要]
遊客賞鷹　灰面鷲不賣帳	2篇 [瀏覽式摘要] [重點式摘要]

■國際新聞

Internet

Browsing (1)



總統保證絕不加稅八大結論振興財經

[1] [2] [3] [4] [5] [6] [7] [8] [下一篇]

資料來源：http://www.chinatimes.com.tw/news/papers/online/focus/c89a1510.htm

尹乃菁、張瑞昌／台北報導行政院財經小組擴大會議昨天在總統府舉行，做成八大結論，包括絕不加稅；

水不缺、電不少；

免除金融營業稅，金融業對營運及繳息正常的傳統產業，有財務周轉需要者，自即日起六個月內到期需償還的本金，可展延半年；

為振興傳統產業，政府將協調八大行庫推出四千五百億專案貸款，如有擔保不足的問題，得申請中小企業信保基金提供保證。

陳水扁總統也在致詞時強調，經濟發展是台灣永續的命脈，更是政府施政的優先考量，他保證新政府一定會提供穩定的投資環境，也呼籲朝野政黨領袖握手和解、共商國是。

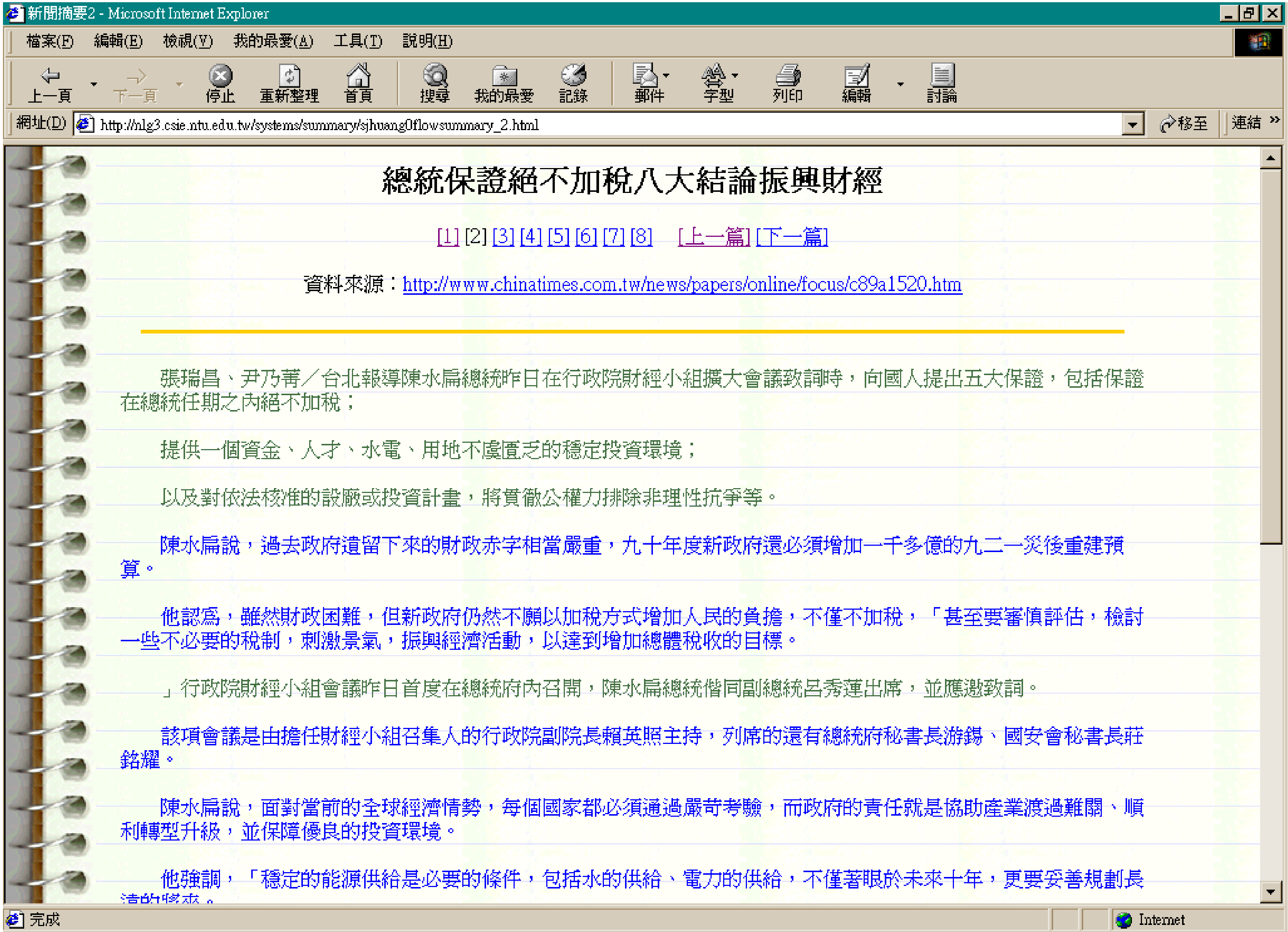
財經情勢低迷、股市重挫，為了尋求改善國內投資環境、提振景氣，確立標本兼治的政策方向，行政院昨天在總統府召開財經小組擴大會議。

新政府對這項會議高度重視，總統、副總統親自出席、全程與會，政府財經首長、四大工商團體領導人在兩個半小時的會議後，做成八項政策結論。

四大工商團體負責人對新政策都表示高度滿意與肯定。

陳水扁總統致詞時提出五大保證：絕不加稅；

Browsing (2)



總統保證絕不加稅八大結論振興財經

[1] [2] [3] [4] [5] [6] [7] [8] [上一篇] [下一篇]

資料來源：http://www.chinatimes.com.tw/news/papers/online/focus/c89a1520.htm

張瑞昌、尹乃菁／台北報導陳水扁總統昨日在行政院財經小組擴大會議致詞時，向國人提出五大保證，包括保證在總統任期之內絕不加稅；

提供一個資金、人才、水電、用地不虞匱乏的穩定投資環境；

以及對依法核准的設廠或投資計畫，將貫徹公權力排除非理性抗爭等。

陳水扁說，過去政府遺留下來的財政赤字相當嚴重，九十年度新政府還必須增加一千多億的九二一災後重建預算。

他認為，雖然財政困難，但新政府仍然不願以加稅方式增加人民的負擔，不僅不加稅，「甚至要審慎評估，檢討一些不必要的稅制，刺激景氣，振興經濟活動，以達到增加總體稅收的目標。

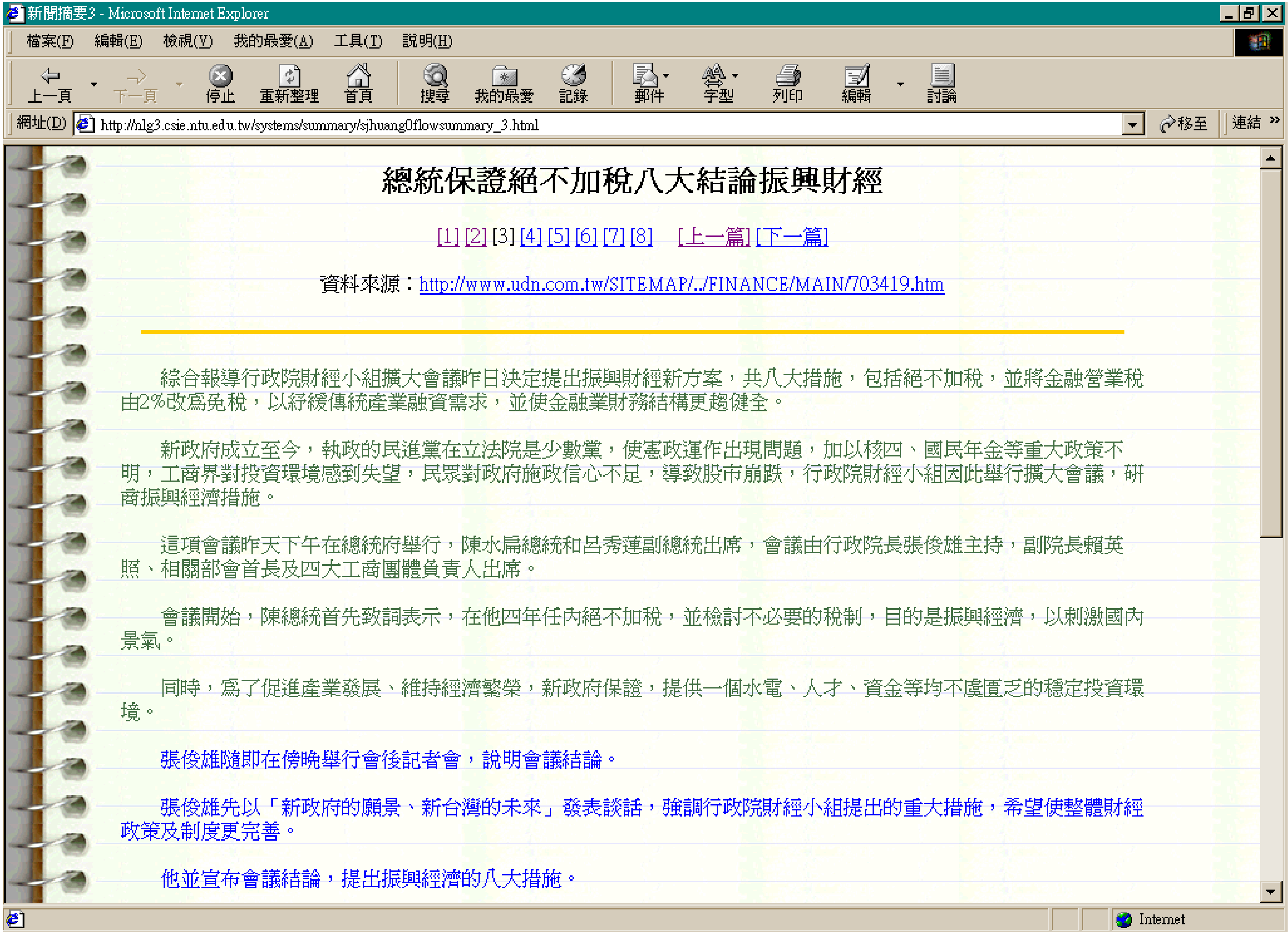
」行政院財經小組會議昨日首度在總統府內召開，陳水扁總統偕同副總統呂秀蓮出席，並應邀致詞。

該項會議是由擔任財經小組召集人的行政院副院長賴英照主持，列席的還有總統府秘書長游錫堃、國安會秘書長莊銘耀。

陳水扁說，面對當前的全球經濟情勢，每個國家都必須通過嚴苛考驗，而政府的責任就是協助產業渡過難關、順利轉型升級，並保障優良的投資環境。

他強調，「穩定的能源供給是必要的條件，包括水的供給、電力的供給，不僅著眼於未來十年，更要妥善規劃長遠的將來。

Browsing (3)



總統保證絕不加稅八大結論振興財經

[1] [2] [3] [4] [5] [6] [7] [8] [上一篇] [下一篇]

資料來源：http://www.udn.com.tw/SITEMAP/../FINANCE/MAIN/703419.htm

綜合報導行政院財經小組擴大會議昨日決定提出振興財經新方案，共八大措施，包括絕不加稅，並將金融營業稅由2%改為免稅，以紓緩傳統產業融資需求，並使金融業財務結構更趨健全。

新政府成立至今，執政的民進黨在立法院是少數黨，使憲政運作出現問題，加以核四、國民年金等重大政策不明，工商界對投資環境感到失望，民眾對政府施政信心不足，導致股市崩跌，行政院財經小組因此舉行擴大會議，研商振興經濟措施。

這項會議昨天下午在總統府舉行，陳水扁總統和呂秀蓮副總統出席，會議由行政院長張俊雄主持，副院長賴英照、相關部會首長及四大工商團體負責人出席。

會議開始，陳總統首先致詞表示，在他四年任內絕不加稅，並檢討不必要的稅制，目的是振興經濟，以刺激國內景氣。

同時，為了促進產業發展、維持經濟繁榮，新政府保證，提供一個水電、人才、資金等均不虞匱乏的穩定投資環境。

張俊雄隨即在傍晚舉行會後記者會，說明會議結論。

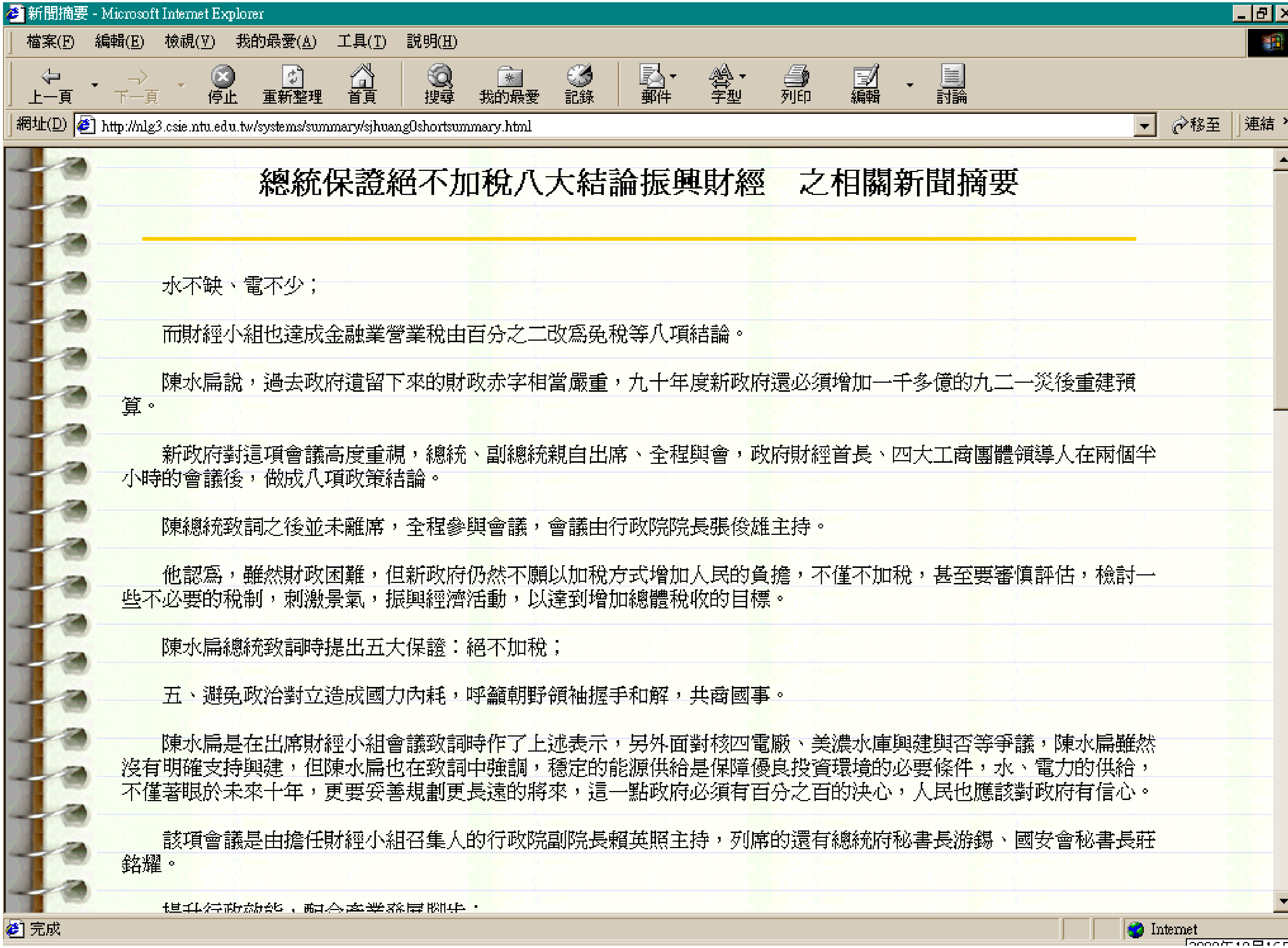
張俊雄先以「新政府的願景、新台灣的未來」發表談話，強調行政院財經小組提出的重大措施，希望使整體財經政策及制度更完善。

他並宣布會議結論，提出振興經濟的八大措施。

Focusing Model

- For each event, a reporter records a news story from his own viewpoint.
- Those MUs that are similar in a specific event are common focuses of different reporters.
- For readability, the original sentences that cover the MUs are selected.
- For each set of similar MUs, only the longest sentence is displayed.
- The display order of the selected sentences is determined by relative position in the original news articles.

Focusing Model



總統保證絕不加稅八大結論振興財經　之相關新聞摘要

水不缺、電不少；

而財經小組也達成金融業營業稅由百分之二改爲免稅等八項結論。

陳水扁說，過去政府遺留下來的財政赤字相當嚴重，九十年度新政府還必須增加一千多億的九二一災後重建預算。

新政府對這項會議高度重視，總統、副總統親自出席、全程與會，政府財經首長、四大工商團體領導人在兩個半小時的會議後，做成八項政策結論。

陳總統致詞之後並未離席，全程參與會議，會議由行政院院長張俊雄主持。

他認爲，雖然財政困難，但新政府仍然不願以加稅方式增加人民的負擔，不僅不加稅，甚至要審慎評估，檢討一些不必要的稅制，刺激景氣，振興經濟活動，以達到增加總體稅收的目標。

陳水扁總統致詞時提出五大保證：絕不加稅；

五、避免政治對立造成國力內耗，呼籲朝野領袖握手和解，共商國事。

陳水扁是在出席財經小組會議致詞時作了上述表示，另外面對核四電廠、美濃水庫興建與否等爭議，陳水扁雖然沒有明確支持興建，但陳水扁也在致詞中強調，穩定的能源供給是保障優良投資環境的必要條件，水、電力的供給，不僅著眼於未來十年，更要妥善規劃更長遠的將來，這一點政府必須有百分之百的決心，人民也應該對政府有信心。

該項會議是由擔任財經小組召集人的行政院副院長賴英照主持，列席的還有總統府秘書長游錫、國安會秘書長莊銘耀。

Experiments and Evaluation

- measurements
  - the document reduction rate
  - the reading-time reduction rate
  - the information carried
- The higher the document reduction rate is, the more time the reader may save, but the higher possibility the important information may be lost

Reduction Rates for Focusing Summarization

Event Name	Doc Len	Sum Len	Sum/Doc	Reduction%
1. military service	7658	2402	0.3137	68.63%
2. construction permit	4182	1226	0.2932	70.68%
3. landslide in Shan Jr	5491	1823	0.3320	66.80%
4. Bush's sons	6186	924	0.1494	85.06%
5. Typhoon Babis	4068	1460	0.3589	64.11%
6. stabilization fund	8434	2243	0.2659	73.41%
7. theft of Dr Sun Yat-sen's calligraphy	4576	1524	0.3330	66.70%
8. interest rate of the Central Bank	4578	1690	0.3692	63.08%
9. the resignation issue of the Cabinet	4980	1368	0.2747	72.53%
Average	50153	14660	0.2923	70.77%

Reduction Rates for Browsing Summarization

Event Name	Doc Len	Sum Len	Sum/Doc	Reduction%
1. military service	7658	2716	0.3547	64.53%
2. construction permit	4182	2916	0.6973	30.27%
3. landslide in Shan Jr	5491	2946	0.5365	46.35%
4. Bush's sons	6186	5098	0.8241	17.59%
5. Typhoon Babis	4068	2270	0.5580	44.20%
6. stabilization fund	8434	4299	0.5097	49.03%
7. theft of Dr Sun Yat-sen's calligraphy	4576	2840	0.6206	37.94%
8. interest rate of the Central Bank	4578	2682	0.5858	41.42%
9. the resignation issue of the Cabinet	4980	3190	0.6406	35.94%
Average	50153	28957	0.5774	42.26%

Ratio of Summary to Full Article in Browsing Summarization

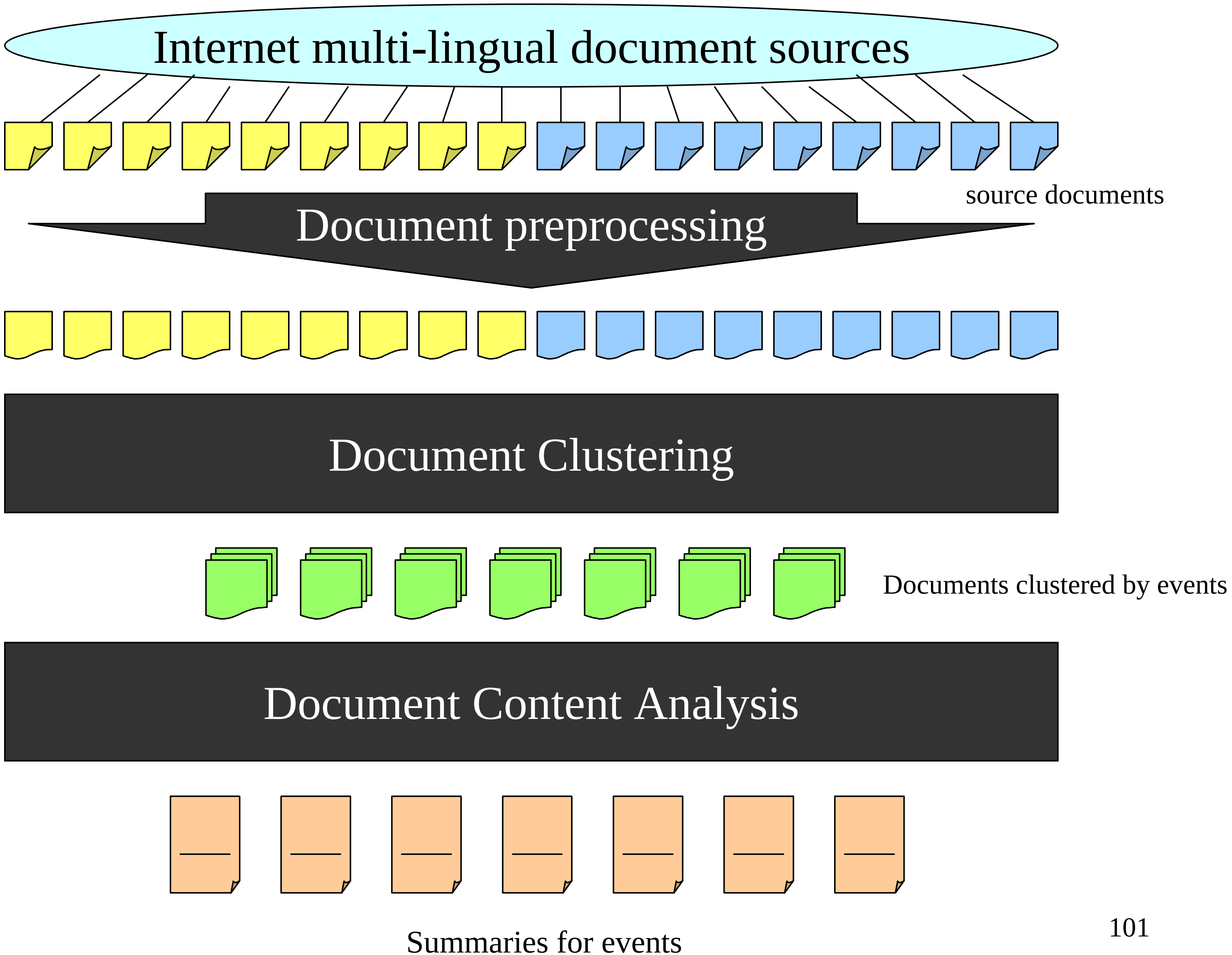
Article\Event	1	2	3	4	5	6	7	8	9
1	100%	100%	100%	100%	100%	100%	100%	100%	100%
2	12%	84%	68%	71%	56%	39%	27%	29%	67%
3	32%	36%	68%	77%	0%	7%	49%	24%	50%
4	24%	47%	10%	79%		51%			24%
5	12%		0%			17%			
6	0%		9%						

Assessors' Evaluation

Event Name	Document Reduction Rate	Question-Answering Correct Rate	Reading-Time Reduction Rate
1. military service	64.53%	100%	45.24%
2. construction permit	30.27%	33.33%	33.54%
3. landslide in Shan Jr	46.35%	80%	10.28%
4. Bush's sons	17.59%	100%	36.49%
5. Typhoon Babis	44.20%	100%	35.10%
6. stabilization fund	49.03%	100%	18.49%
Average	43.79%	88.46%	30.86%

Issues in Multilingual Summarization

- Translation among news stories in different languages
- Idiosyncrasy among languages
- Implicit information in news reports
- User preference

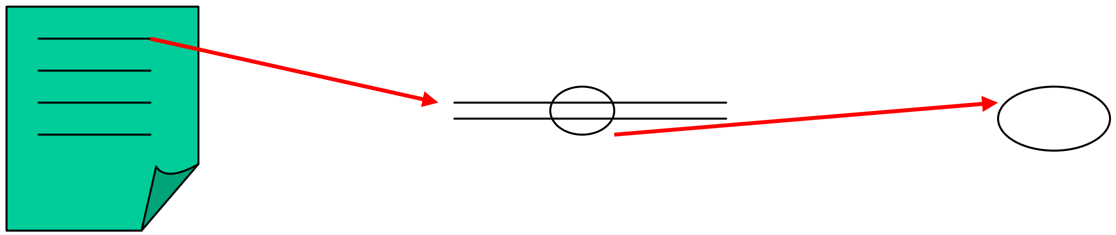

Internet multi-lingual document sources
source documents
Document preprocessing
Document Clustering
Documents clustered by events
Document Content Analysis
Summaries for events

Issues

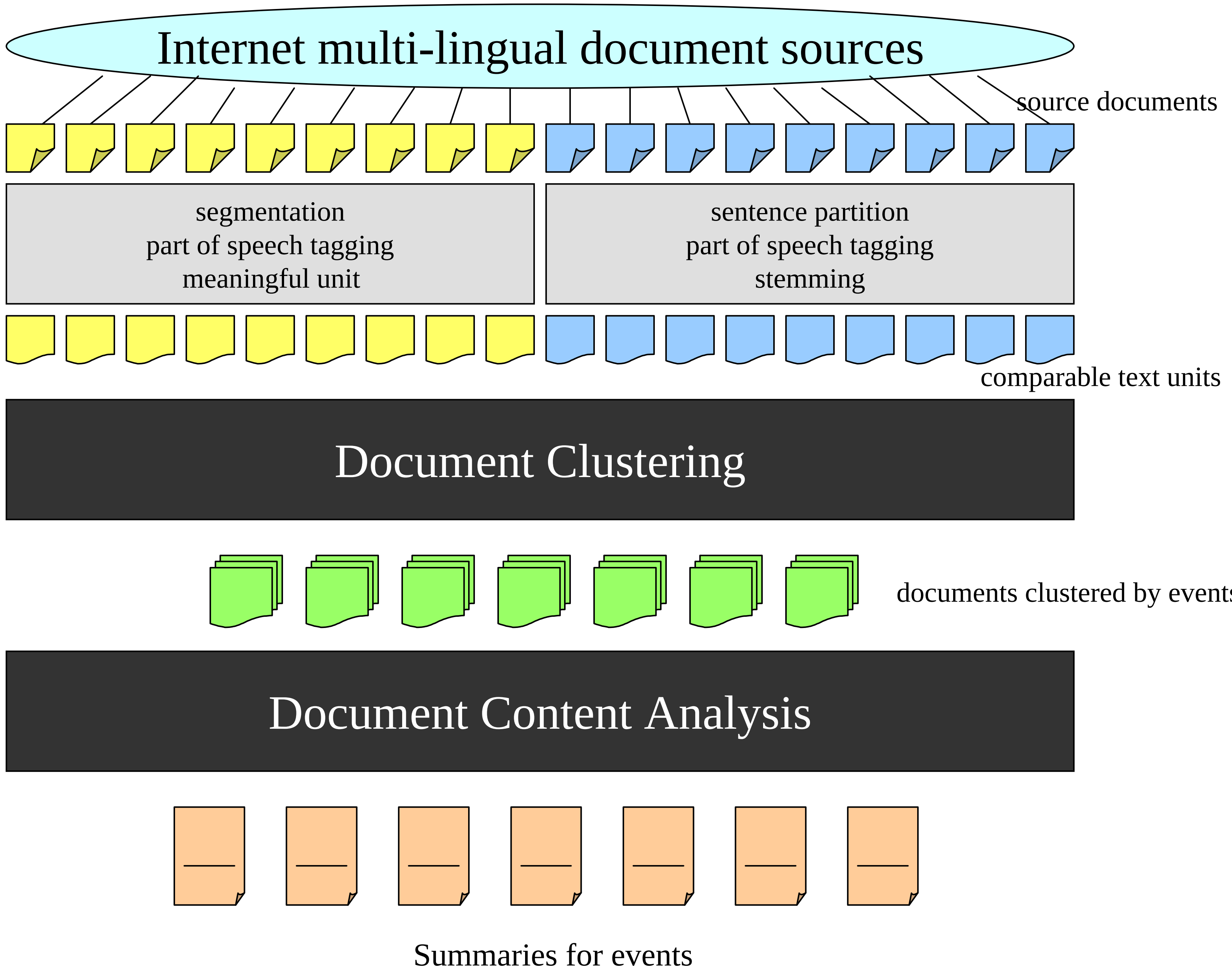
- How to represent Chinese/English documents?
- How to measure the similarity between Chinese/English representations?
  - word/phrase level
  - sentence level
  - document level
- Visualization

Document Preprocessing

- Comparable Units

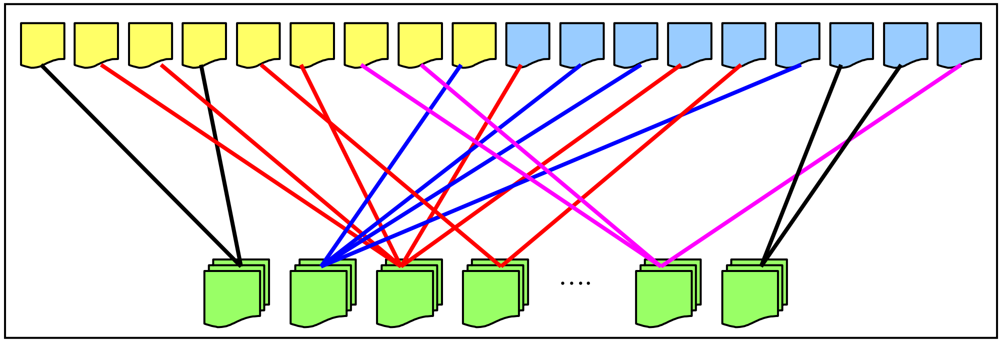


	document	passage	word
Chinese	document	meaningful unit	word (segmentation)
English	document	sentence	word

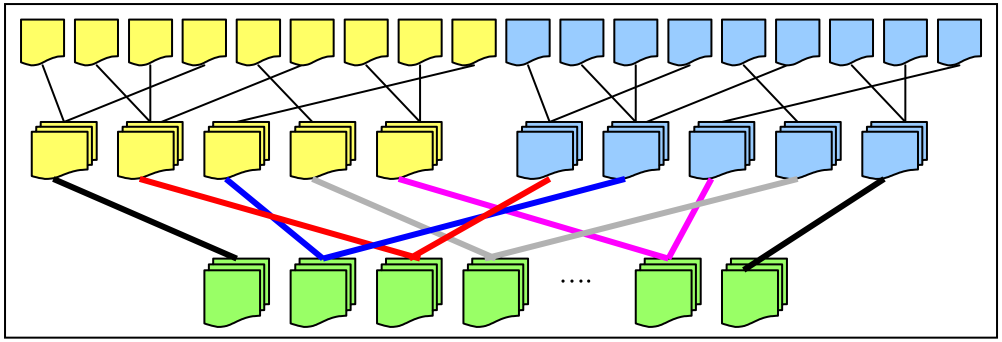

Internet multi-lingual document sources
source documents
segmentation
part of speech tagging
meaningful unit
sentence partition
part of speech tagging
stemming
comparable text units
Document Clustering
documents clustered by events
Document Content Analysis
Summaries for events

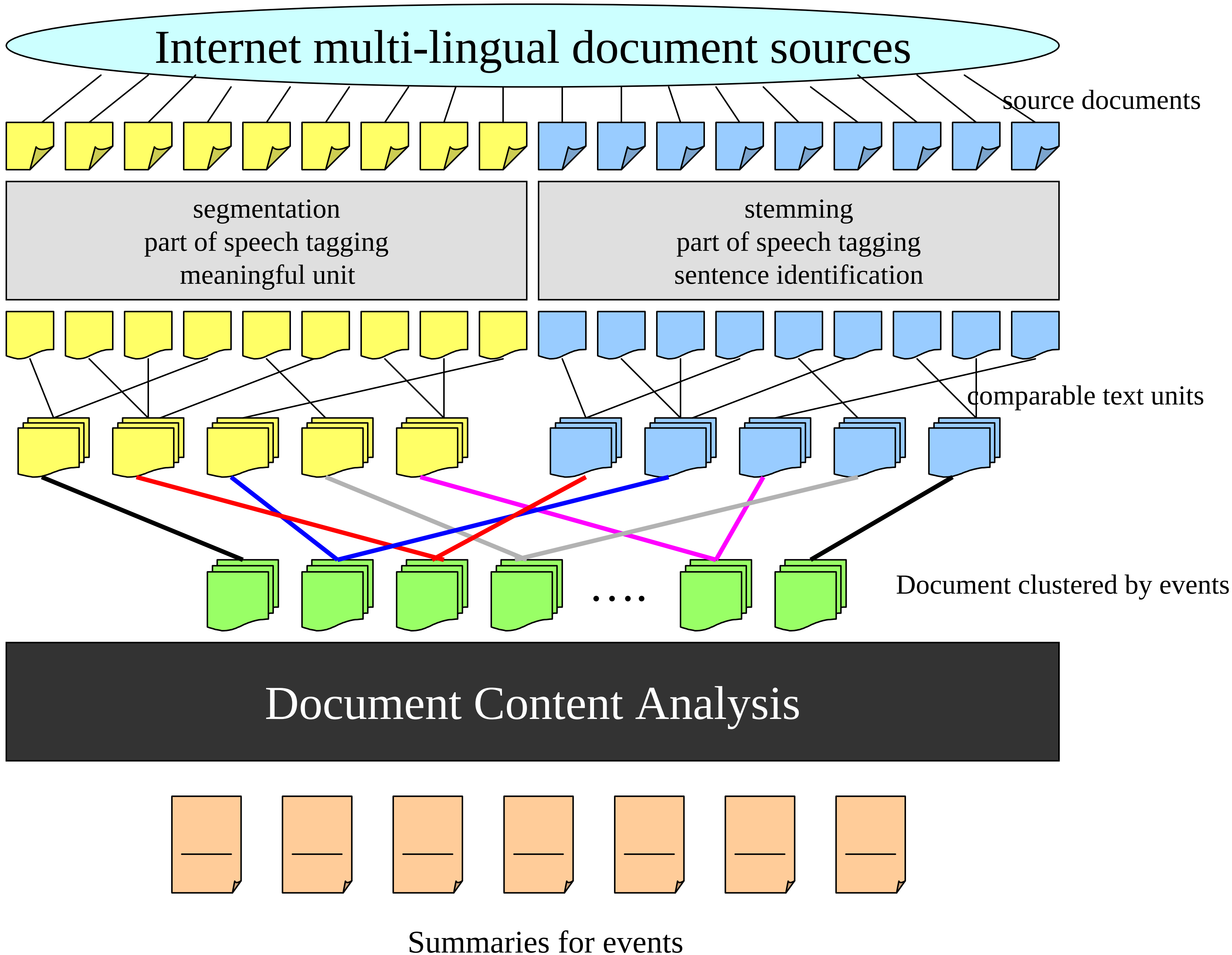
Document Clustering

Alternative 1: Clustering English and Chinese documents TOGETHER

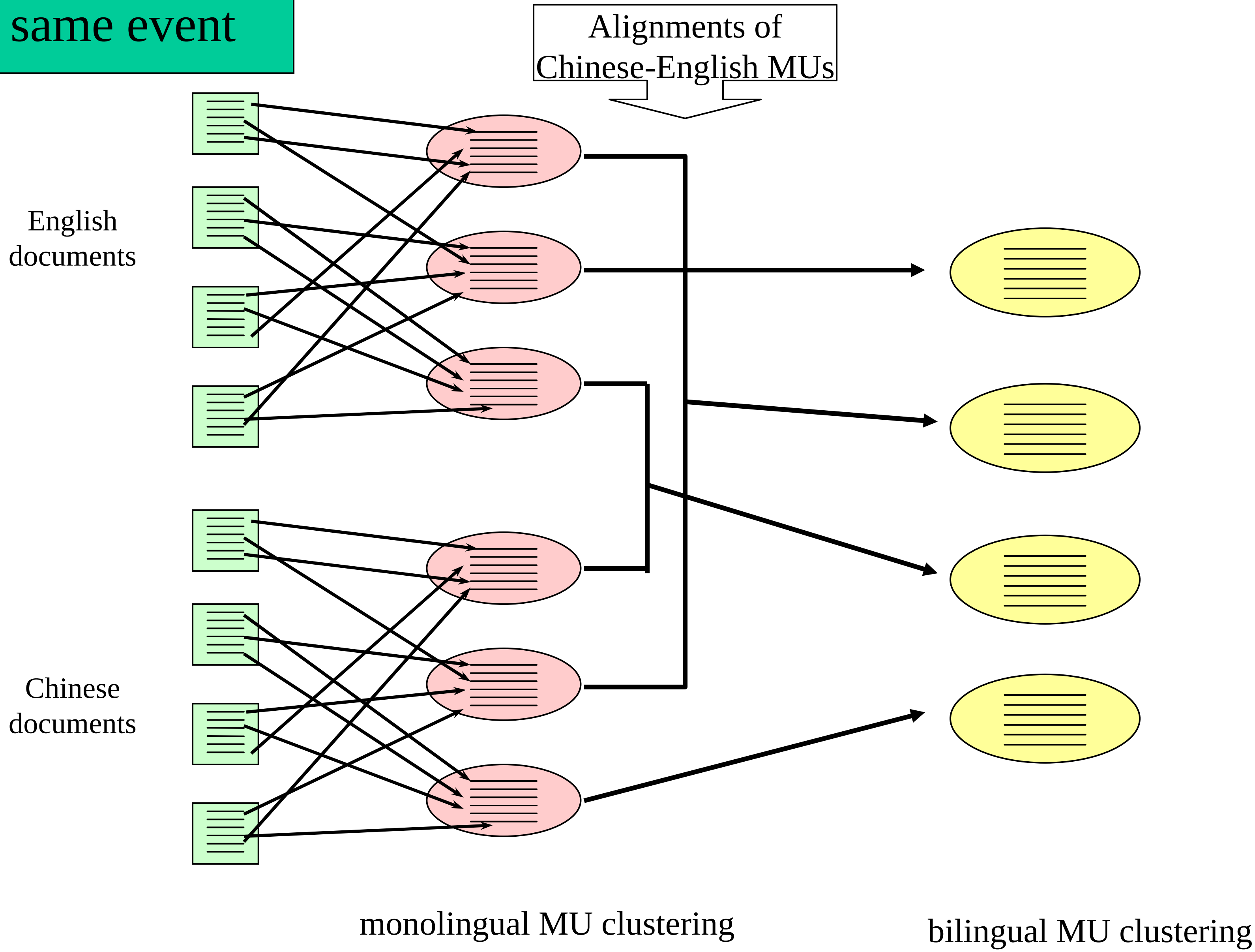


Alternative 2: Clustering English and Chinese documents SEPARATELY and merging clusters



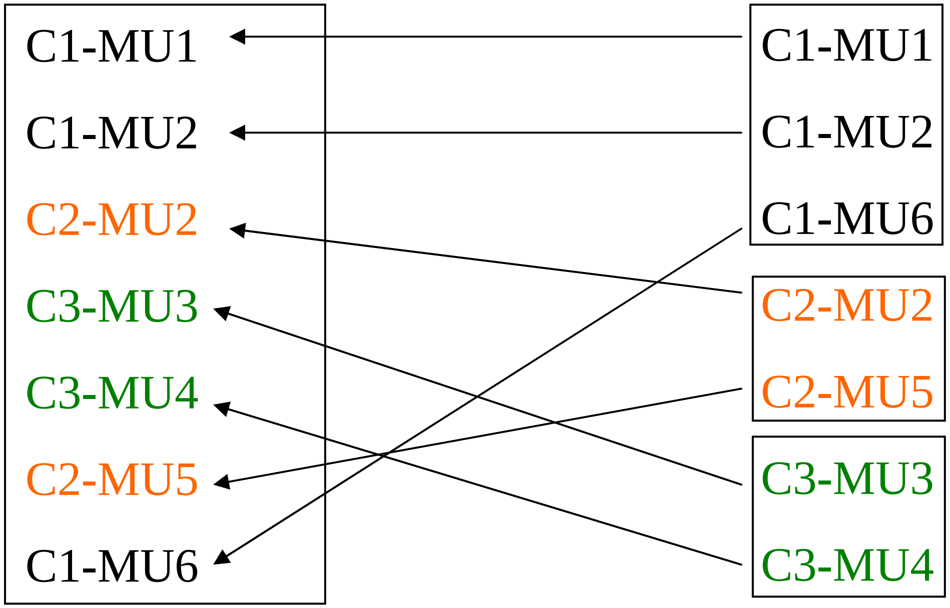

Internet multi-lingual document sources
source documents
segmentation
part of speech tagging
meaningful unit
stemming
part of speech tagging
sentence identification
comparable text units
Document clustered by events
....
Document Content Analysis
Summaries for events

same event


Alignments of
Chinese-English MUs
English
documents
Chinese
documents
monolingual MU clustering
bilingual MU clustering

Visualization

- Focusing summary



Visualization

- focusing summarization

Prefer Chinese	Prefer English
C1-MU1	E2-MU1
C1-MU2	E1-MU2
C2-MU2	E1-MU3
C3-MU3	E5-MU3
C3-MU4	E6-MU4
C2-MU5	E2-MU5
C1-MU6	C2-MU1
E2-MU1	C3-MU2
E1-MU2	C3-MU3
E1-MU3	C1-MU6

Visualization

- browsing

中文1-1
中文1-2
中文1-3
中文1-4
中文1-5
中文1-6

中文2-1
中文2-2
中文2-3
中文2-4
中文2-5

中文3-1
中文3-2
中文3-3

中文4-1
中文4-2

英文1-1
英文1-2
英文1-3
英文1-4
英文1-5

英文2-1
英文2-2
英文2-3

....

英文n-1
英文n-2

Summary

- Topic Detection and Tracking
  - Topic Detection
- Summarization
  - Multiple Document Summarization
  - Multi-Lingual Multi-Document Summarization