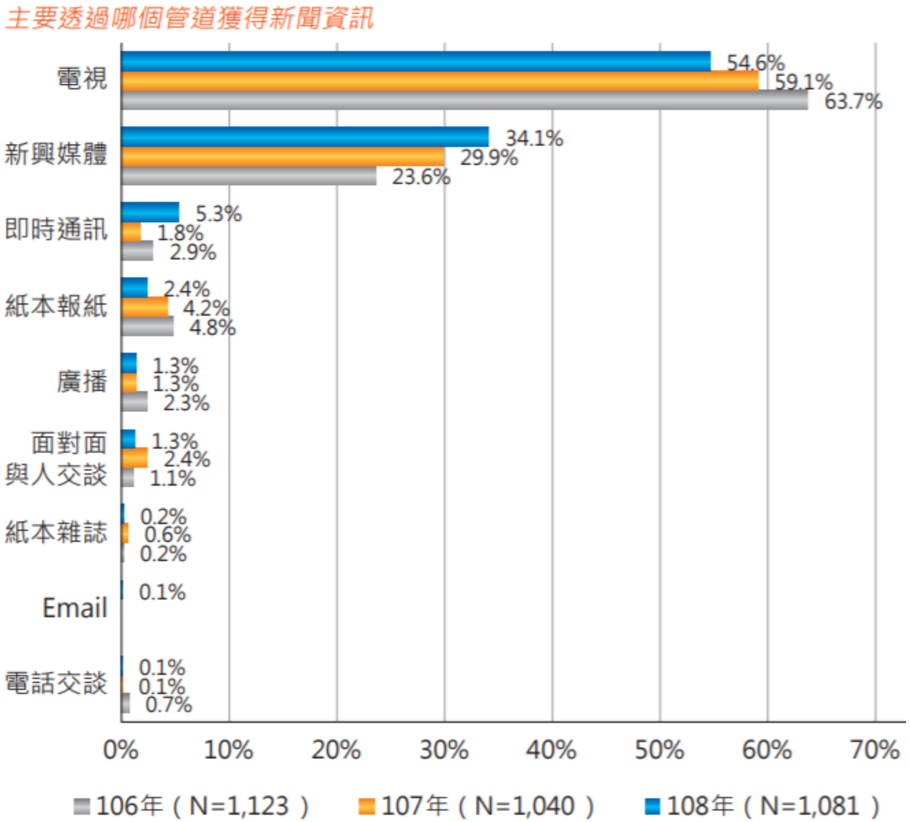
Automatic Text Summarization using closed-Pattern-Infused Edit Similarity (PIESim)
閉合模式融合於編輯相似度之自動文本摘要

台大統計碩士學位學程 陳育婷
指導教授：蔡政安 教授 許聞廉 教授

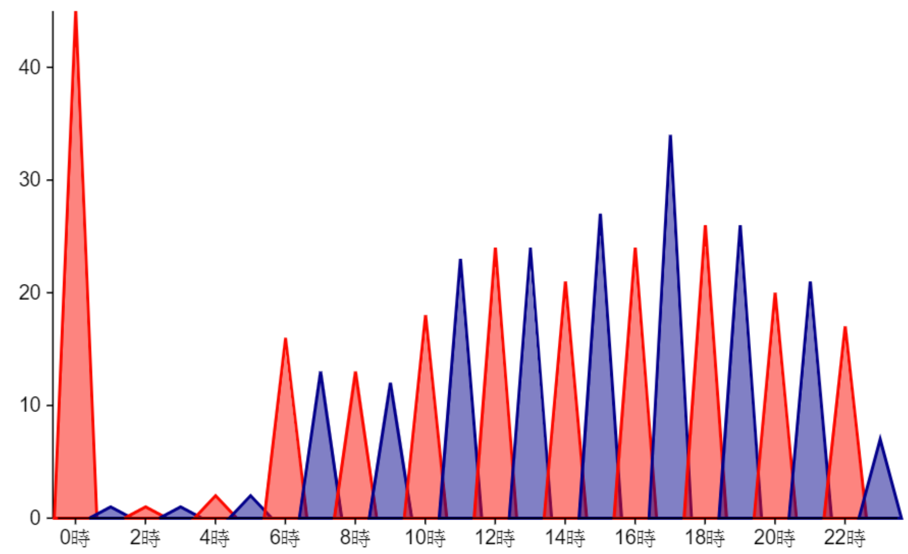
2020/ 07/ 27

Table of Content

Introduction & Aims



108/6/1- 7/31 National Communication Commission (NCC) Digital Convergence Survey with 95% C.I.

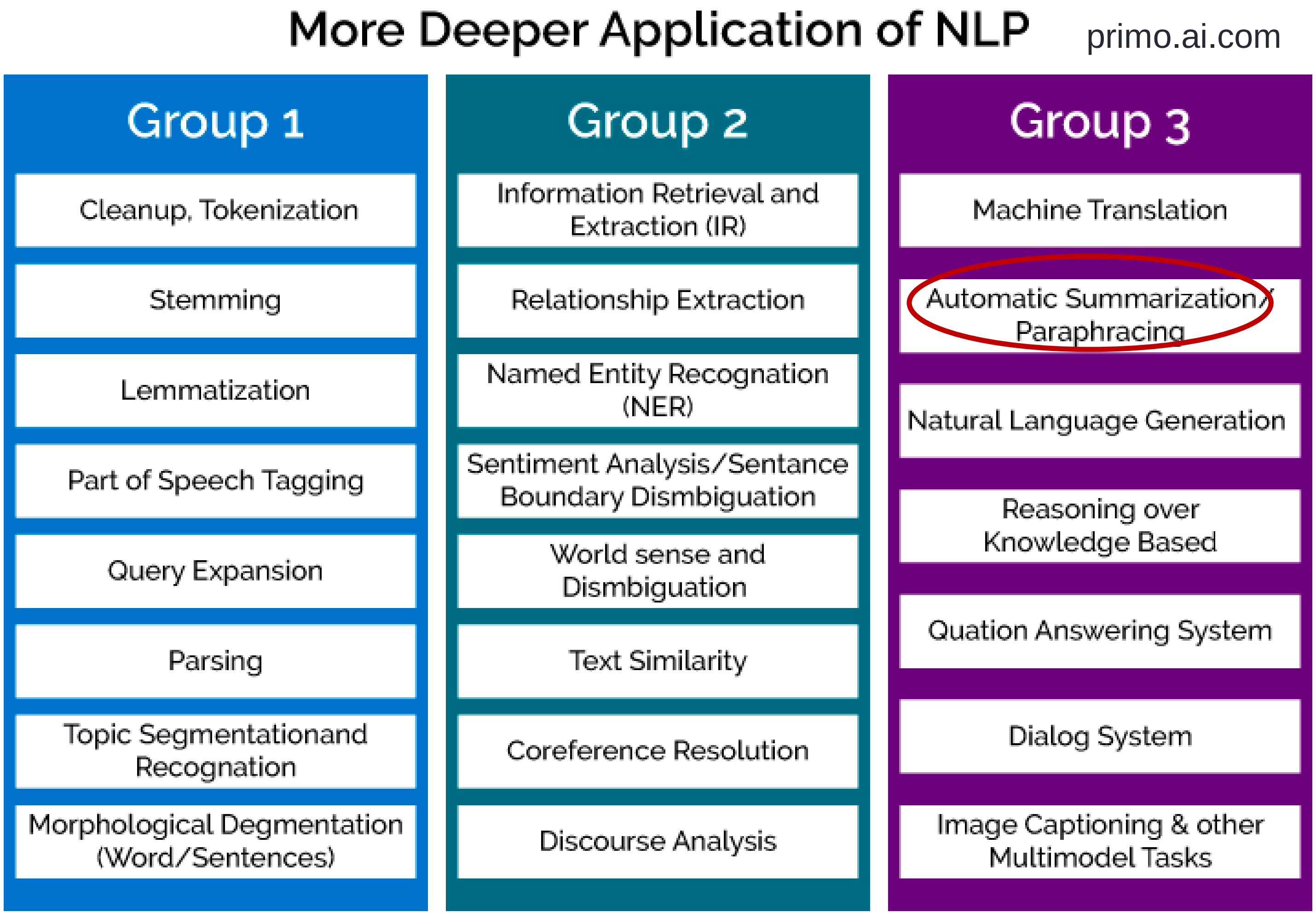


Number of streamed News on 2019/04/07 of Apple Daily News

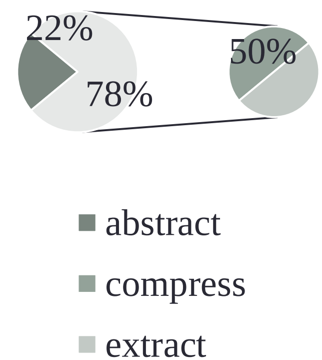
- ✓ Preliminary Goal : A precise and revisable automatic text summarization (ATS) system can ease the burden of news publishers and benefit all kinds of human

Introduction & Aims

- ✓ Preliminary Goal : design an extractive, precise, and revisable summarization system with compression techniques



primo.ai.com



Sources of Summaries

- A. Aries and W. K. Hidouci, "Automatic text summarization: What has been done and what has to be done,"

Introduction & Aims

What's Important ?

No systems can cover all aspects. Considering from general and query aspects is more useful in real life

→ Incorporate Query Aspect

Use of Enormous Data

Most non DL or ML approaches do not use information outside the document

→ Structure can consider new info.

Vector-based Represen-tation

Term-based : Lack Semantic Information

Deep learning : Hard to revise

Not conform to human process

→ Non-vector rep. with semantic info.

✓ Goal : design an extractive, precise, non-vector-based , revisable, query & new-info.-considered ATS system with compression techniques

Introduction & Aims

precise

Performance is the best among large categories of ATS systems on Chinese and English dataset.

non-vector-based & revisable

Non-vector based but preserves semantic information; thus can be intuitively revised

query & new-info.-considered

Structure is easy to absorb new info. We propose a novel memory similarity criterion to learn from training data

innovative

The first to combine pattern and edit distance in ATS and to propose a new PIESim similarity measure

Literature Review – Genres of ATS

Stati-stical

Pros: simple, explanation and revision

Cons: many use word-independent features

Graph-based

Pros: exploit shared information between units, more coherent

Cons: constructed from the target document only

Lingu-istic

Pros : usually more accurate

Cons : restricted to languages and domains, more laboring work

ML

Pros : More complex and automatic

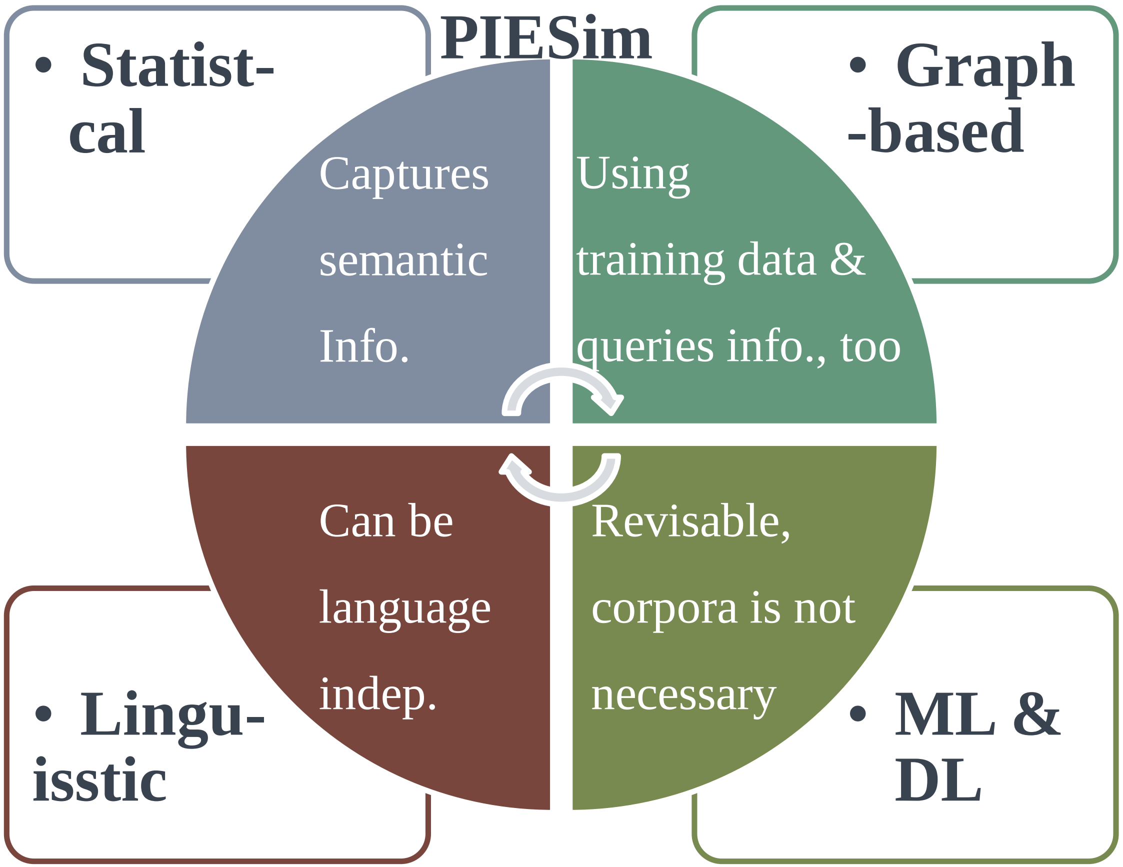
Cons : Needs corpora to learn the rules, model itself is not easy to revise.

DL

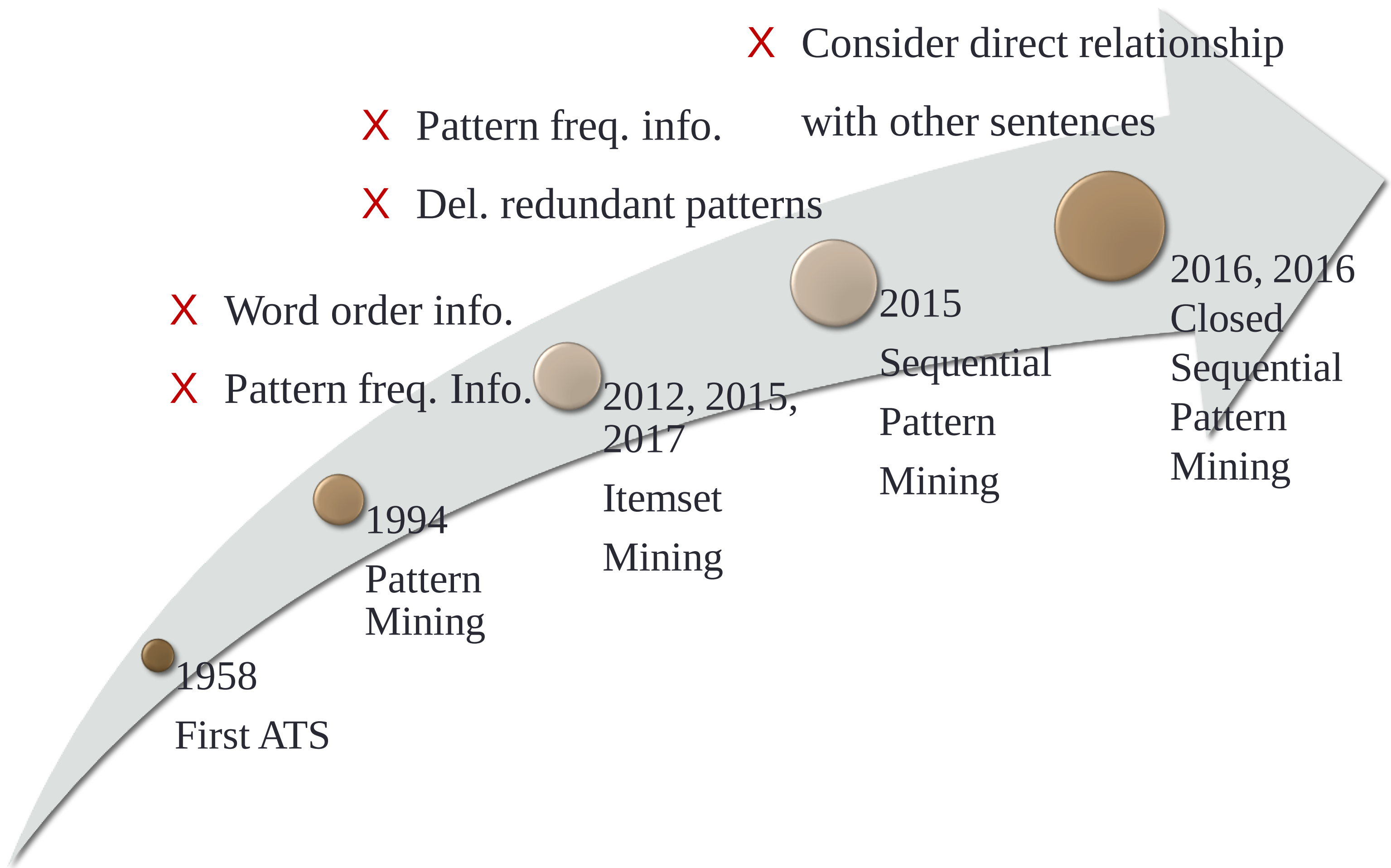
Pros : Fully automatic. Perform well when corpora is enough

Cons : Explanation and revision itself is not easy to revise. Often non grammatically correct.

Literature Review – Genres of ATS



Literature Review - Pattern based ATS



Literature Review – Edit Dist. in ATS

Preprocessing Tool

- graph-based to cluster similar-phrase node or string-based to pre-filter sent.s

Sole Sim. Measurement Pattern-based weighted Edit Distance

- consider contextual similarity, and with intuitive explanation and revision ability

Similarity Measurement

- graph-based or term-based weighted edit dist.

Background Study – Categorization of ATS

Type		Description	
Extractive	Abstractive	Sentence selection	Rephrasing, reinterpreting concepts
Single document	Multiple documents	Summarization of single document	Summarization of multiple document
Generic	Query-Focused	Preserve general important parts	Consider only user preferences
Supervised	Unsupervised	Use training data	Not using training data

Background Study – Sequential Pattern Mining

A subfield of pattern mining. Discover interesting (sequential) patterns in a (sequential) database

- $I : I = \{i_1; i_2; \ldots. i_m\}$, i_k:$item\ k$. A set of all items.
- $itemset\ X : X \subseteq I$.
- $sequence\ S : S = \langle I_1, I_2, \ldots, I_n \rangle$. An ordered itemset list such that $I_k \subseteq I$.
- sequence $S_a = \langle A_1, A_2, \ldots, A_m \rangle$ *is contained/ subsequence* in/of sequence $S_b = \langle B_1, B_2, \ldots, B_n \rangle$:

  $S_a \sqsubseteq S_b$ if $\exists 1 \leq i_1 \leq i_2 \leq \cdots \leq n\ s.t.\ \forall A_k \in S_a,\ A_k \subseteq B_{i_k}$
- $sequential\ database\ SDB : SDB = \langle S_1, S_2, \ldots, S_n \rangle$. A list of sequences.
- $support\ sup(S_a) : sup(S_a) = |S | S_a \sqsubseteq S \wedge S \in SDB|$. Number of sequences in a SDB containing S_a

Background Study – Sequential Pattern Mining

SID	Sequence	Sequential Patterns (minsup = 2)	Closed Sequential Patterns (minsup = 2)
1	$\langle\{a\},\{b,c\},\{d,e\}\rangle$	$\langle\{c\}\rangle, \langle\{d\}\rangle, \langle\{e\}\rangle$ $\langle\{c\},\{d\}\rangle, \langle\{c\},\{e\}\rangle, \langle\{d\},\{e\}\rangle$ $\langle\{d,e\}\rangle$	$\langle\{c\},\{d\}\rangle, \langle\{c\},\{e\}\rangle, \langle\{d\},\{e\}\rangle$ $\{d,e\}$
2	$\langle\{c\},\{d,e\}\rangle$		
3	$\langle\{e,f\}\rangle$		

A Sequential Database Example with $I = \{a, b, c, d, e, f\}$

1. Order Matters :

   $\langle\{b\},\{d,e\}\rangle \sqsubseteq S_1 = \langle\{a\},\{b,c\},\{d,e\}\rangle$ but $\langle\{d,e\},\{b\}\rangle \not\sqsubseteq S_1 = \langle\{a\},\{b,c\},\{d,e\}\rangle$

2. Sequential Patterns/ Frequent Subsequences FS:

   $FS = \{S_{sub} \mid |S_{sub} \sqsubseteq S \wedge S \in SDB| \geq minsup\}$.

3. Closed Sequential Patterns CS :

   $CS = \{S_a \mid S_a \in FS \wedge \nexists S_b \in FS \ st. S_a \sqsubset S_b \wedge \sup(S_a) = \sup(S_b)\}$.

   lossless representation, largest subsequences common to sets of sequences

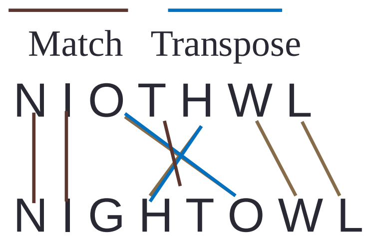
Background Study – Edit Distance

- Distance Measure between two units (usually strings) : counting number of edit operations needed to transform one string to another
- Examples : Substitute Insert Transpose Delete

Levenshtein Distance	Damerau–Levenshtein Distance	Longest Common Subsequence Distance
N I O T H W L	N I O T H W L	N I O T H W L
N I G H T O W L	N I G H T O W L	N I G H T O W L
$\frac{4}{8} = 0.5$	$\frac{3}{8} = 0.375$	$\frac{5}{8} = 0.625$
Substitution Insertion Deletion	Substitution Insertion Deletion Transposition (2 adjacent units)	Insertion Deletion

Background Study – Jaro Similarity

Transposition



$$1-\frac{1}{3}\left(\frac{7}{7}+\frac{7}{8}+\frac{7-1}{7}\right)=0.2898$$

$$sim_j=\frac{1}{3}\left(\frac{|m|}{|s_1|}+\frac{|m|}{|s_2|}+\frac{|m|-\frac{|t|}{2}}{|m|}\right), dist_j=1-sim_j$$

$$m=\langle (t_{1i},t_{2j}) | t_{1i}=t_{2j} \wedge |p_{1i}-p_{2j}| \leq \left\lfloor \frac{\max(|s_1|,|s_2|)}{2} \right\rfloor -1 \wedge i \notin \cup_{k=1}^{i-1} k \rangle$$

$$m_{order} = \langle (u_{1k},u_{2k}) | (u_{1k},u_{2k}) \in m \wedge \exists 1 \leq i_1 \leq i_2 \leq \cdots \leq l \ s.t.\ \forall (u_{1k},u_{2k}), u_{1k}=t_{1i_k}, u_{2k}=t_{2i_k} \rangle$$

$$t=\langle (u_{1k},u_{2k}) | (u_{1k},u_{2k}) \in m_{order} \wedge u_{1k} \neq u_{2k} \rangle$$

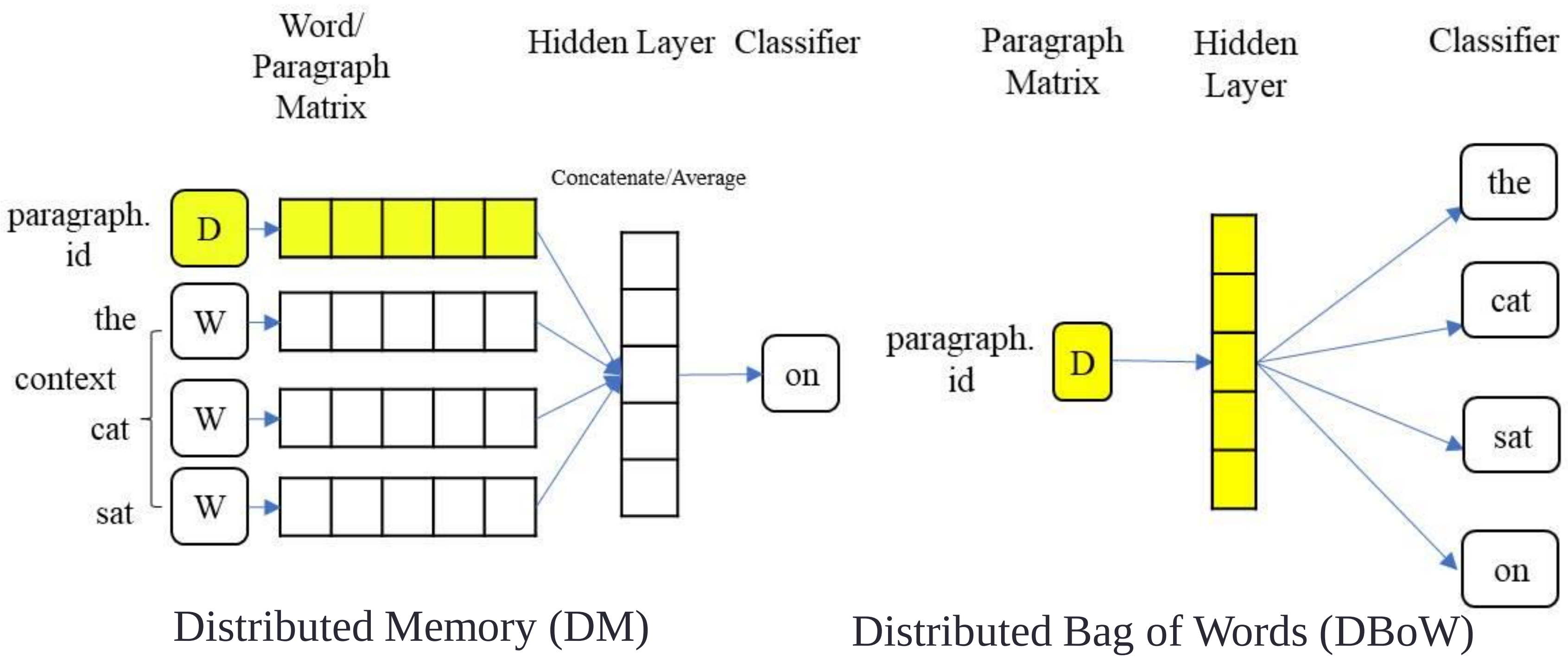
$p_{li}: position\ of\ t_{li}\ in\ s_l$

- Perform Well

-W. Cohen, P. Ravikumar, and S. Fienberg,“A comparison of string metrics for matching names and records”

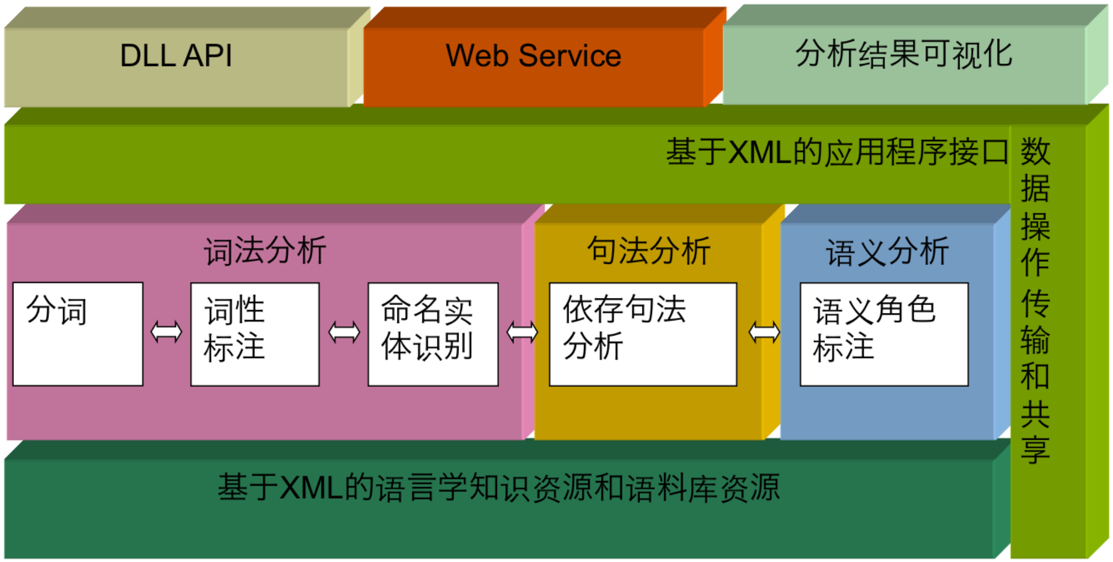
-G. Recchia and M. M. Louwerse,“A Comparison of String Similarity Measures for Toponym Matching”

Background Study – Paragraph Embedding

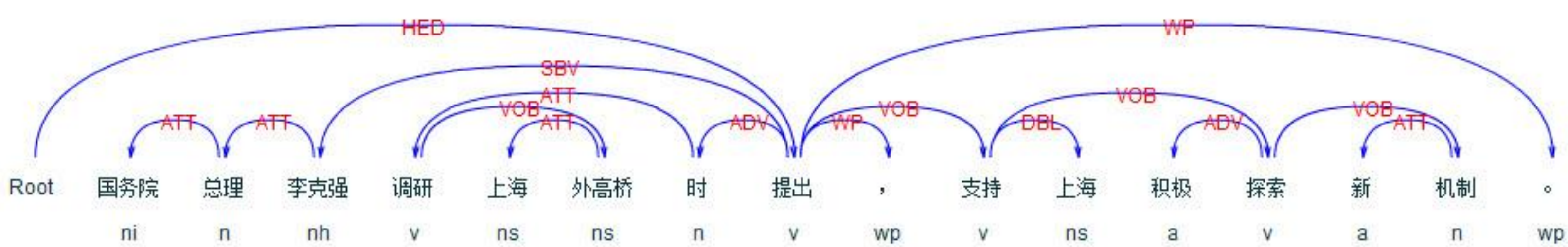


- ✓ Learn contextual information by predicting words from their context in the model
- ✓ Indirectly captures semantics of the paragraph

Background Study – Tools & Knowledge for Sent. Compression



Framework of Language Technology Platform (LTP) provided by HIT-SCIR (LTP official page)



Example of a Dependency Parsing Structure (LTP official page)

Background Study – Rouge-n

- Full name : Recall-Oriented Understudy for Gisting Evaluation
- To compute the number of units (n-grams) in both the system's summary and the reference one and calculate the recall

✓ $Rouge\text{ - }n = \dfrac{\sum_{S_{ri} \epsilon Summ_{ref}} \sum_{S_{rij} \in S_{ri}} |\{ngram|ngram \in S_{rij} \wedge ngram \epsilon S_{si}\}|}{\sum_{S_{ri} \epsilon Summ_{ref}} \sum_{S_{rij} \in S_{ri}} |\{ngram|ngram \in S_{rij}\}|}$

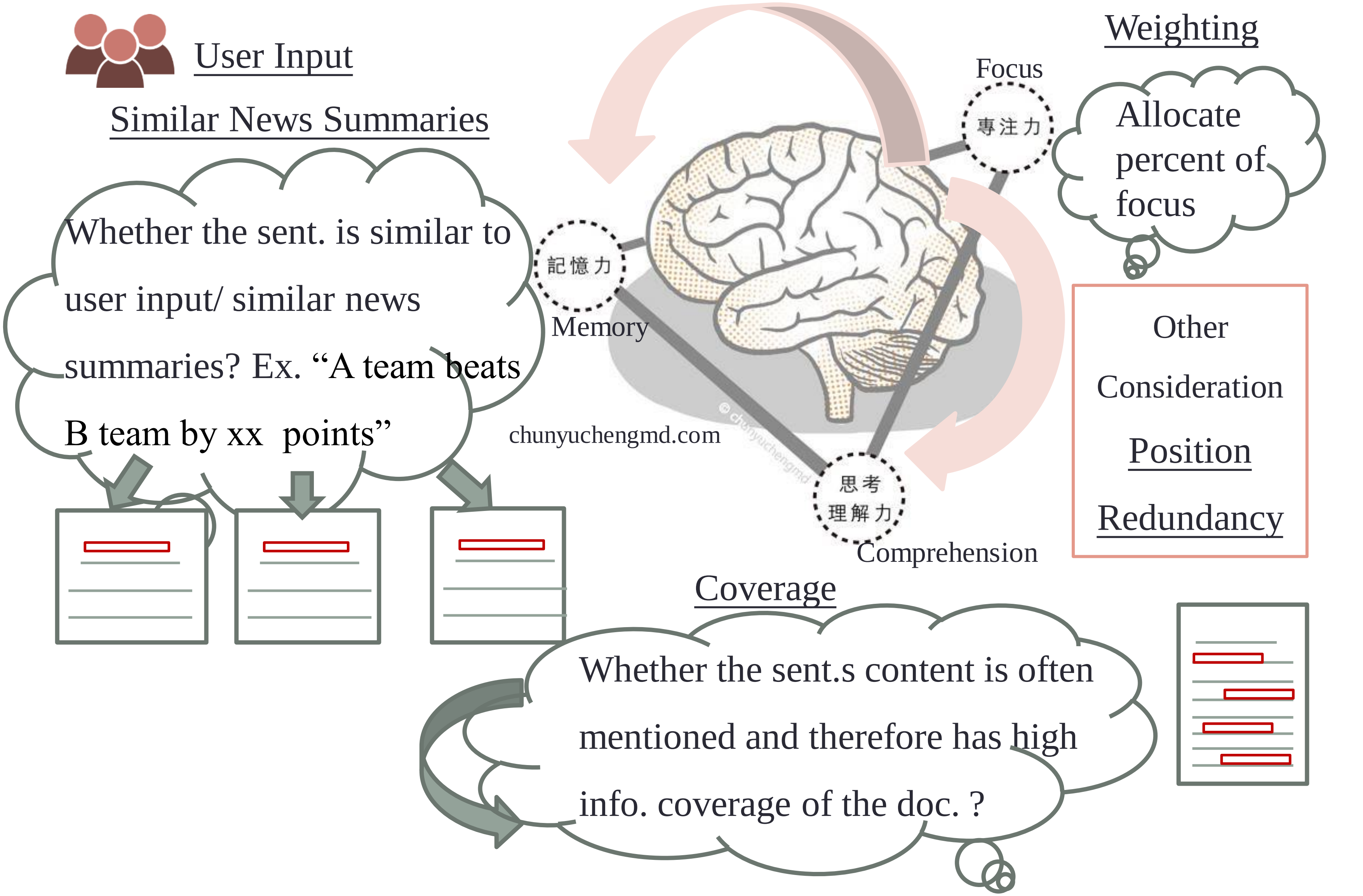
$Summ_{ref}$: set of reference summaries, $Summ_{sys}$: set of system summaries

S_{rij}: jth summary in ith summary set in $Summ_{ref}$, S_{si}: ith summary in $Summ_{sys}$

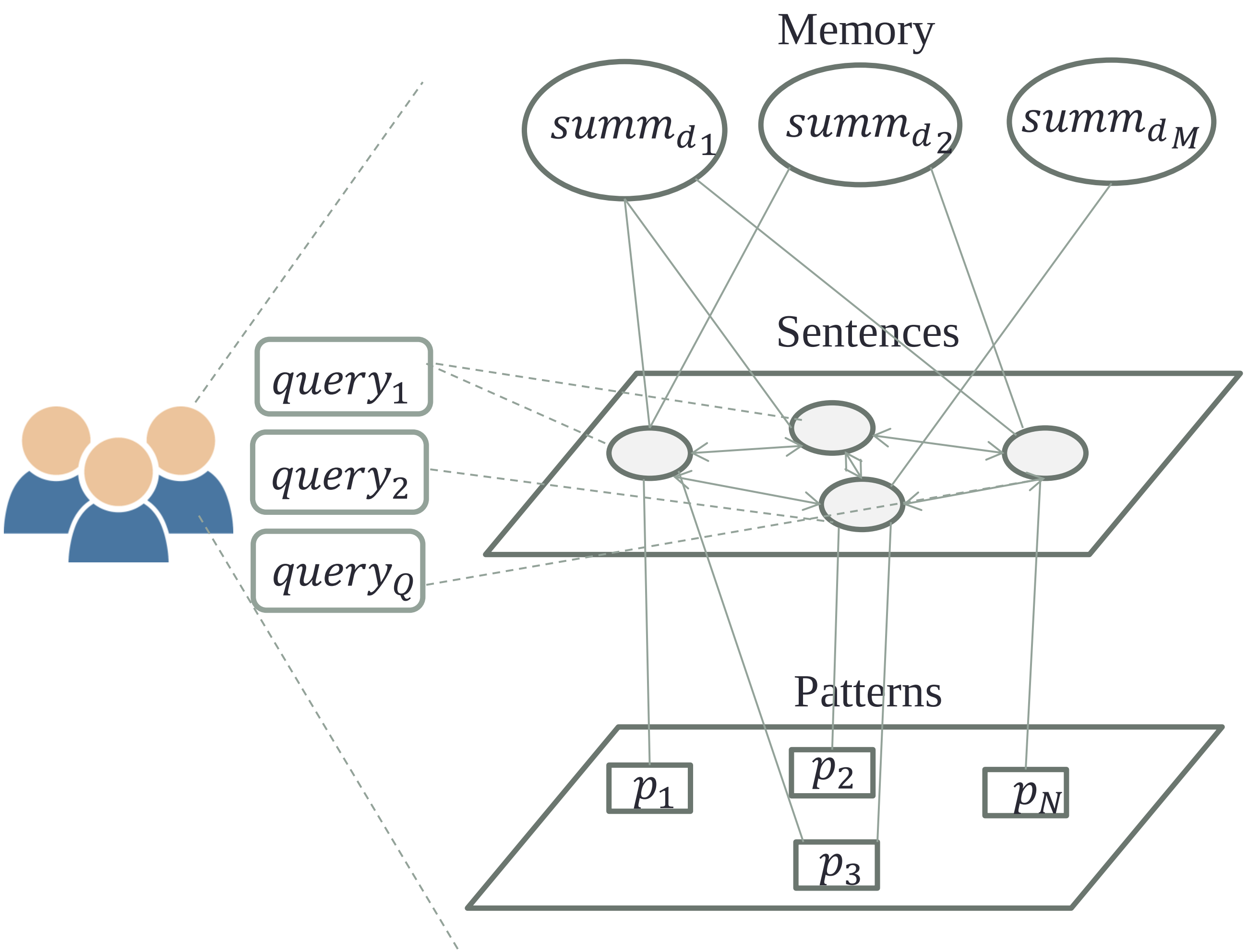
✓ *Precision - n* : replacing denominator with total counts of n-grams in system summaries

✓ F1 : $\dfrac{1}{\frac{1}{2}(Rouge - n + Precision - n)}$

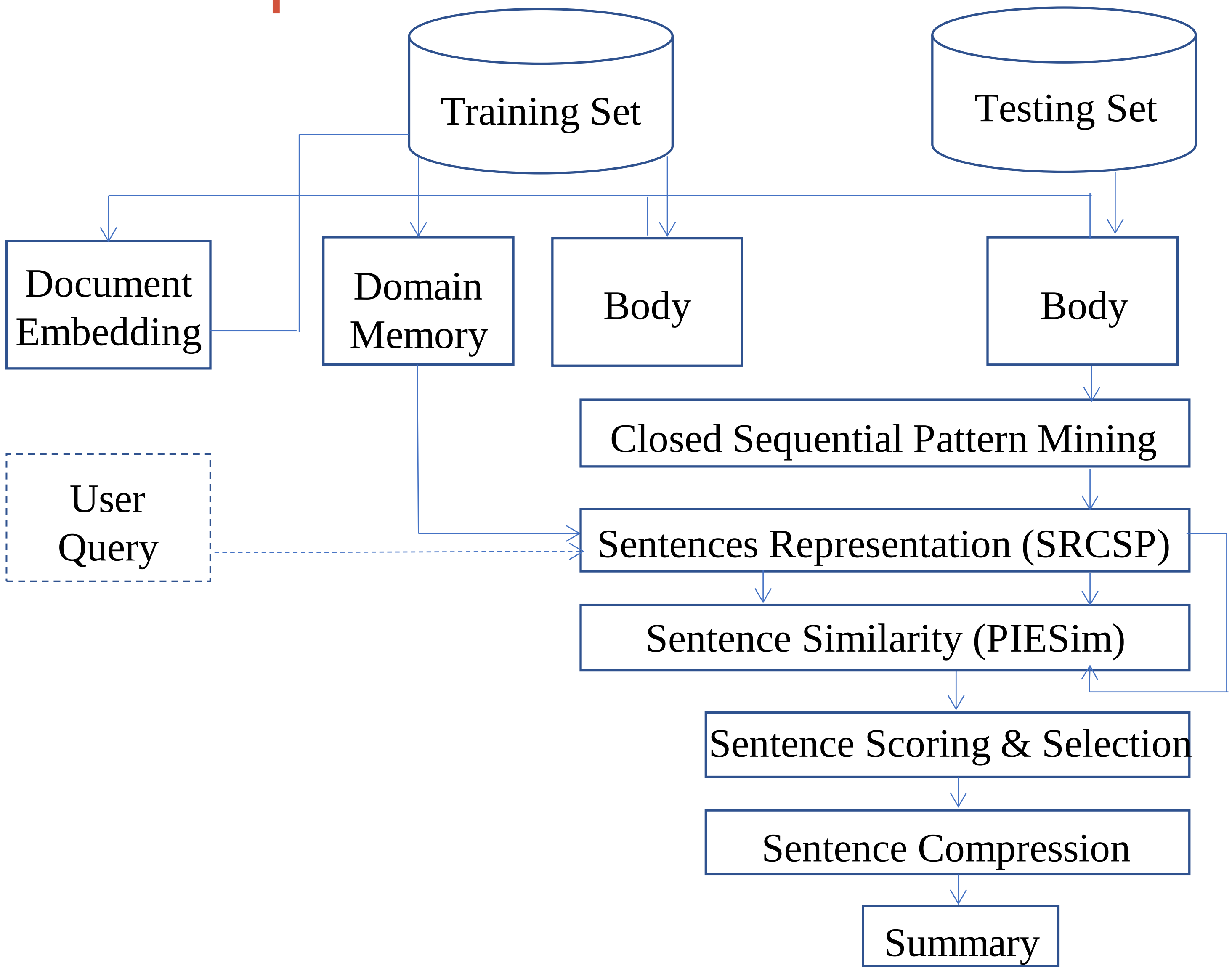
The Proposed PIESim Model – Analogy to Human



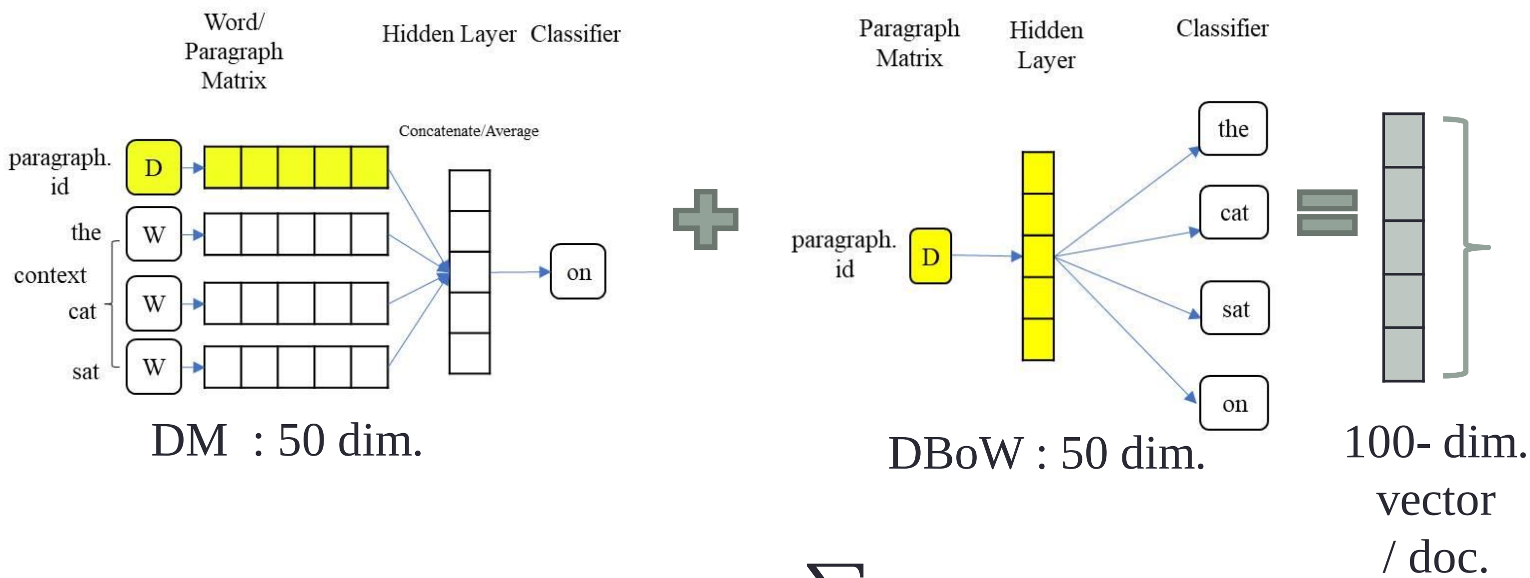
The Proposed PIESim Model – Structure



The Proposed PIESim Model – WorkFlow



The Proposed PIESim Model – Memory Retrieval



$$memory_d = \{s_i | i \in argmax_{D' \subset D, |D'|=10} \sum_{j \in D \setminus d} cos(embed_d, embed_j)\}$$

$$cos(embed_d, embed_j) = \frac{embed_d \cdot embed_j}{|embed_d||embed_j|}$$

s_i: ith $document's$ $summary$, $embed_j$: jth $document's$ $embedding$,

$D = \{i | i \in 1,2, \ldots, |training\ doc.s|\}$

The Proposed PIESim Model – Sent. Rep. using Closed Sequential Patterns (SRCSP)

Definition 1 Let $p_i = \langle t_1, t_2, \ldots t_m \rangle$ be a closed sequential pattern with $sup(p_i)$; t_j is *jth* term and *m* is count of terms in p_i. A "pattern-weight-pair set" of a pattern is

$$pw(p_i) = \{(t_1, w), (t_2, w), \ldots (t_m, w)\},\ w = sup(\ p_i)\ , t_j \in p_i$$

Definition 2 Let $pw(p_i) = \{(t_1, w_i), (t_2, w_i), \ldots (t_m, w_i)\}$ and $pw(p_j) = \{(s_1, w_j), (s_2, w_j), \ldots (s_n, w_j)\}$ be two pattern-weight-pair sets, the interchangeable composition operation $\oplus$ of them is formulized below:

$$pw(p_i) \oplus pw(p_j) =$$

$$\{(t_i\ , w_i + w_j) | \{(t_i\ , w_i + w_j) | (t_i\ , w_i) \in\ pw(p_i) \wedge (s_j\ , w_j) \in pw(p_j) \wedge\ t_i =\ s_j\}$$

$$\cup\ \ \{(\ t_i\ , w_i) | (t_i\ , w_i) \in\ pw(p_i), t_i\ \notin \bigcup_{j=1}^{n} s_j\}$$

$$\cup\ \ \{(s_j\ , w_j) | (s_j\ , w_j) \in\ pw(p_j), s_j\ \notin \bigcup_{i=1}^{m} t_i\}$$

The Proposed PIESim Model - SRCSP

Definition 3 A sentence representation of a sentence sequence $s = \langle t_1, t_2, \dots, t_m \rangle$ is the function of ordered composition operation of pattern-weight pairs from patterns contained in the sentence.

term-based rep.

$$s_{pw} = \{pw(p_i) | p_i \sqsubseteq s, p_i \in CS_d\}$$

CS_d : closed sequential patterns of a document d

$$sentrep(s) = pw_{s1} \oplus pw_{s2} \oplus pw_{sk}, \qquad pw_{si} \in s_{pw}$$

$$sentrep_{term}(s) = \langle (a_i, w_i) | s_{pw} \neq \emptyset \wedge (a_i, w_i) \in sentrep(s) \wedge \exists 1 \leq i_1 \leq i_2 \leq \cdots \leq m \; s.t. \forall (a_k, w_k), a_k = t_{ik} \wedge a_i \notin \bigcup_{j=1}^{i-1} a_j \rangle \cup \langle (a_i, 1) | s_{pw} = \emptyset \wedge a_i \in s \wedge \exists 1 \leq i_1 \leq i_2 \leq \cdots \leq n \; s.t. \forall a_k, a_k = t_{ik} \rangle$$

char-based rep.

$$s_c = \langle c_{11}, \dots, c_{1|t_1|}, c_{10}, c_{21}, \dots c_{m-1|t_{m-1}|}, c_{(m-1)0}, c_{m1} \dots c_{m|t_m|} \rangle$$

t_i : term i in s, c_{ij} : character j in term i (include space)

$$f(c_{ij}) = \begin{cases} w_i, j \neq 0 \\ \frac{1}{8}, j = 0 \end{cases}, \quad c_{ij} \in s_c, w_i \in \langle w_k | (t_k, w_k) \in sentrep_{term}(s) \rangle$$

$$sentrep_{char.}(s) = \langle \left(s_{ij_k}, f(s_{ij})_k \right) | s_{ij} \in s_c \rangle$$

The Proposed PIESim Model – SRCSP Examples

Def. 1 p_i = *{read, book}* with *support 2* -> $pw(p_i)$ = *{(read, 2), (book, 2)}*

Def. 2 $pw(p_j)$ = *{(buy, 3),(book, 3)}* -> $pw(p_i) \oplus pw(p_j)$ = *{(book, 5), (buy, 3), (read, 2)}*

Def. 3	**Original Sentence**	**Contained CS**	**SRCSP**
English	⟨I, usally, buy, a, book, from, the, bookstore, and, read, the, book, at, night⟩	{(buy, 3), (book, 3)} {(read, 2), (book, 2)}	⟨(buy, 3), (book, 5), (read, 2)⟩
Chinese	⟨我, 通常, 從, 書店, 買, 書本, 然後, 在, 晚上, 讀, 書本⟩	{(買,3),(書本,3)}, {(讀,2),(書本,2) }	⟨(買,3),(‘ ‘,$\frac{1}{8}$) (書,5), (本,5), (‘ ‘, $\frac{1}{8}$),(讀,2)⟩

- ***English or lang.s with few char.s:***, matching in unit of character will be misleading
- ***Chinese or lang.s with many char.s***: 50000- 100000 characters. Character match actually implies semantic relation in many cases. Ex. “腳踏車(bicycle)” & “汽車(car)” Will be inappropriate sometimes, gain is larger than loss of info. In our experiments

The Proposed PIESim Model – PIESim Measure

$$s_{rep1} = sentrep_{term(character)}(s_1),\; s_{rep2} = sentrep_{term(character)}(s_2)$$

$$PIESim(s_1, s_2) = \frac{1}{3}\left(\frac{m_w}{\sum_{\{w_i | (t_i, w_i) \in s_{rep1}\}} w_k} + \frac{m_w}{\sum_{\{w_j | (t_j, w_j) \in s_{rep2}\}} w_k} + \frac{m_w - \frac{|t|}{2}}{m_w}\right)$$

$$m = \langle (c_{1i}, c_{2j}) | c_{1i} = c_{2j} \;\wedge\; |p_{1i} - p_{2j}| \le \left\lfloor \frac{\max(|s_{rep1}|, |s_{rep2}|)}{2} \right\rfloor - 1 \wedge i \notin \bigcup_{k=1}^{i-1} k \rangle$$

$$m_w = \sum_{\substack{\{(w_{1_i}, w_{2_j}) | (c_{1i}, c_{2j}) \in m \wedge (c_{1i}, w_{1i}) \in s_{rep1} \wedge \\ (c_{2j}, w_{2j}) \in s_{rep2}\}}} \frac{w_{1_i} + w_{2_j}}{2}$$

$$m_{order} = \left\langle \begin{matrix} (u_{1k}, u_{2k}) | \\ (u_{1k}, u_{2k}) \in m \wedge \exists 1 \le i_1 \le i_2 \le \cdots \le l \\ s.t. \\ \forall (u_{1k}, u_{2k}), u_{1k} = c_{1i_k}, u_{2k} = c_{2i_k} \end{matrix} \right\rangle$$

$$t = \langle (u_{1k}, u_{2k}) | (u_{1k}, u_{2k}) \in m_{order} \wedge u_{1k} \ne u_{2k} \rangle$$

c_{li} : *unit i in* s_l, p_{li} : *position of* c_{li} *in* s_l

Sim. Measure	SRCSP of s1 s2	Sim. Computation
PIESim *match*	{(buy, 3), (book, 5), (read, 2)}	$\frac{1}{3}\left(\frac{\frac{5+5}{2} + \frac{2+2}{2} = 7}{10} + \frac{7}{13} + \frac{7-1}{7}\right) \approx 0.7$
Jaro Similarity *transposition*	{(I, 3), (love, 3), (read, 2), (book,5)}	$\frac{1}{3}\left(\frac{2}{4} + \frac{2}{3} + \frac{2-1}{2}\right) \approx 0.55$

The Proposed PIESim Model – Scoring Selection

$$score(s_i) = \{ (1-\beta) * \left[\alpha \ *. \underbrace{\frac{cov(s_i)}{\sum_j cov(s_j)}}_{\text{Coverage}} + (1-\alpha) * \underbrace{\frac{mem(s_i)}{\sum_j mem(s_j)}}_{\text{Memory Sim.}}\right] +$$

$$\underbrace{\beta * \frac{div(s_i)}{\sum_j div(s_j)}}_{\text{Diversity (- Redundancy)}} \} * \underbrace{pos_i}_{\text{Positional Weight}}$$

$$cov(s_i) = \frac{\sum_{s \in D} PIESim(s_i, s) - 1}{|D| - 1}, mem(s_i) = \frac{\sum_{h \in memory_d} PIESim(s_i, h)}{|memory_d|}$$

$$div(s_i) = min(\{1 - PIESim(s_k, s_i) | s_k \in S\})$$

$h \in memory_d, s \in \{sentences\ in\ d\},$

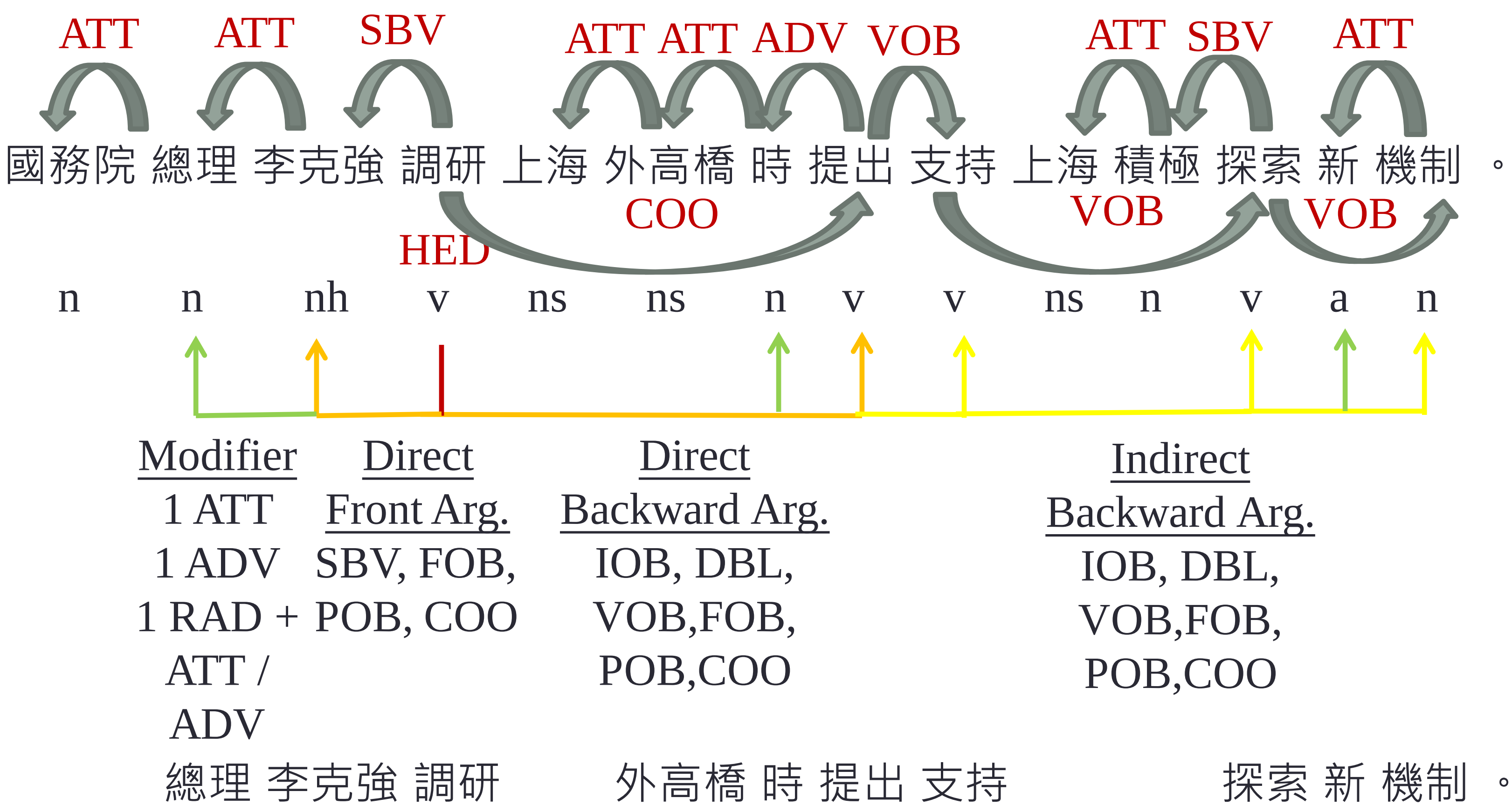
$$pos_i = 1 - \frac{(i-1)}{|\{s|s \in \{sent.s\ in\ document\ d\}\}|},$$

$$pos_i^j = 1 - \frac{(i-1)}{|\left\{s_i^j | s_i^j \in \{\ sent.s\ in\ document\ d_j\}\right\}|}$$

s_i : $ith\ sentence\ of\ d, d$: $document\ d$ S: $summary\ set\ of\ d$

✓ greedily select sent. with highest score iteratively until we reach the limits

The Proposed PIESim Model – Compression (Chinese)



Experiments and Results – Chinese

- 2017 -2018 492997 UDN news ：Select 2 sent. or 1 sent. > 30 char.s

Clean Tags → x Duplicate Headlines → Sent. Split CKIP Tokenizer → Remove Stop Words → $\frac{1}{10}$(49297) Testing

summary stat	med.	std.
word/ hl.	10	2.34
word/ sent. in bd.	7	5.14
char./ word in bd.	2	0.93

Size	**UDN**
Train	443700
Test	49297
Median	~188
Size	**LCSTS**
Train	2407551
Test	1106
Median	~60

System	**PIESim**	**PIESimcomp.**	**ILP**	**Submod.1**	**Submod.2**
Rouge-1 F1	**0.3137**	0.2753	0.2766	0.2647	0.2646
Rouge-2 F1	**0.1836**	0.1520	0.1574	0.1467	0.1467
System	**Luhn**	**SumBasic**	**LSA**	**TextRank**	**LexRank**
Rouge-1 F1	0.2459	0.2706	0.2083	0.2385	0.2083
Rouge-2 F1	0.1352	0.1508	0.1061	0.1330	0.1061
System	**Reduction**	**PatSum**	**Seq2Seq**	**SeqAtt.**	**LCSTS Seq / SeqAtt.**
Rouge-1 F1	0.2385	0.3001	0.1933	0.2607	0.215/ 0.089
Rouge-2 F1	0.1330	0.1775	0.0732	0.1117	0.299/ 0.174

Statistical based

Graph based

Pattern based

DNN

Experiments and Results - English

- Task 2 from benchmark Dataset from DUC 2004 Conference
- 50 clusters with 10 docs. each, No training data – multi. doc.s / unsupervised ATS

✓ Pattern Mining : 10-doc.s / Position Weight : Start from 1 in each doc. / Len. Limit : 665 bytes

Sent. Split (DUC script)

Remove Stop Words (Cornell)

Stem Words (Porter Stemmer)

	Rouge 2			Rouge 4		
	Precision	Recall	F1	Precision	Recall	F1
PIESim	**0.0943**	**0.0953**	**0.0947**	**0.0168**	**0.0170**	**0.0169**
Pat Sum	0.0988	0.0990	0.0986	0.0167	0.0166	0.0166
Freq. Itemset	0.0864	0.0869	0.0866	0.0135	0.0136	0.0135
Weight. Itemset	0.0916	0.0904	0.0909			
Best Peer	0.092	0.091	0.091	0.015	0.015	0.015
Human A	0.088	0.092	0.090	0.010	0.009	0.010
Human B	0.096	0.091	0.092	0.013	0.013	0.013
Human C	0.102	0.094	0.098	0.012	0.011	0.012
Human D	0.106	0.100	0.102	0.010	0.010	0.010
Human E	0.099	0.094	0.097	0.012	0.011	0.012
Human H	0.105	0.101	0.103	0.013	0.012	0.012

Different Settings Analysis – Pattern Usage

*, **, *** represent two-sided-paired-sample t-tests with α = 0.05, 0.01, 0.001 significance level

Tested under 25 combinations of coverage α and redundancy par.s β 0.1, 0.3, 0.5, 0.7, 0.9 on UDN

	Mean Rouge 1 F1 Difference	Mean Rouge 2 F1 Difference	Mean Rouge 1 F1	Mean Rouge 2 F1	Best Rouge 1 F1 / Rouge 2 F2
Use pattern or not – Base : Testing doc. SRCSP w. min. occurr. 2 + Sent.s in Memory's Hl.s					
No Pattern	0.0007	-0.0023	0.2958	0.1678	0.3085/ 0.1756
Train/ Test doc. Patterns	-0.0013***	-0.0010***	0.2937	0.1692	0.3062/ 0.1787
Test doc. Pattern	**base**	**base**	**0.2951**	**0.1702**	**0.3067/0.1789**
Pattern Variant– Base : Testing doc. SRCSP with min. occurrence 2					
Occurrence 3	0.0013	0.0011	0.2965	0.1714	0.3047/ 0.1777
All Occurrences	-0.00017***	- 0.00015***	0.2949	0.1700	0.3067/0.1789
Text Format – Base : Original Testing doc. SRCSP with min. occurrence 2					
Map Ehownet & NER	-0.0111***	-0.0067***	0.2840	0.1634	0.3015/ 0.1756
Map Ehownet	-0.009***	-0.006***	0.2853	0.1632	0.3012/ 0.1744

Different Settings Analysis – PIESim Variant

	Mean Rouge 1 F1 Difference	Mean Rouge 2 F1 Difference	Mean Rouge 1 F1	Mean Rouge 2 F1	Best Rouge 1 F1/ Rouge 2 F1
w. or w,/o. PIESim-weights information (space weight = 1)					
PIESim Weight	**base**	**base**	**0.3012**	**0.1747**	**0.3118/ 0.1821**
No Weight	-0.0060***	-0.0045***	0.2951	0.1702	0.3067/ 0.1789
w. or w,/o. word level information (space weight = 1)					
Not Sep. by Spaces	-0.0065**	-0.0067***	0.2946	0.1680	0.3043/ 0.1742
Matched by words	-0.0200***	-0.0115***	0.2812	0.1632	0.2870/ 0.1666
Space Weight					
1/2	0.0015***	0.0011***	0.3027	0.1758	0.3128/ 0.1829
1/4	0.0020***	0.0015***	0.3032	0.1762	0.3133/ 0.1833
1/8	0.0022***	0.0016***	0.3034	0.1764	0.3137/ 0.1836
0	-0.0077***	-0.0069***	0.2934	0.1678	0.3033/ 0.1739

Different Settings Analysis – Sent. Rep./ Sim. Measure/ Scoring Variant

	Mean Rouge 1 F1 Difference	Mean Rouge 2 F1 Difference	Mean Rouge 1 F1	Mean Rouge 2 F1	Best Rouge 1 F1/ Rouge 2 F1
PIESim	**base**	**base**	**0.3034**	**0.1764**	**0.3137/ 0.1836**
Sent. Rep. & Sim. measurement Variants					
term-freq. infused edit sim.	-0.0088***	-0.0055***	0.2946	0.1709	0.3036/0.1772
tfisf	-0.0165***	-0.0112***	0.2869	0.1652	0.3102/0.1812
sent embedding	-0.0334***	-0.0216***	0.2700	0.1548	0.2905/0.1682
Sent. Scoring					
PageRank	0.0008	0.0011	0.3042	0.1775	0.3121/0.1828

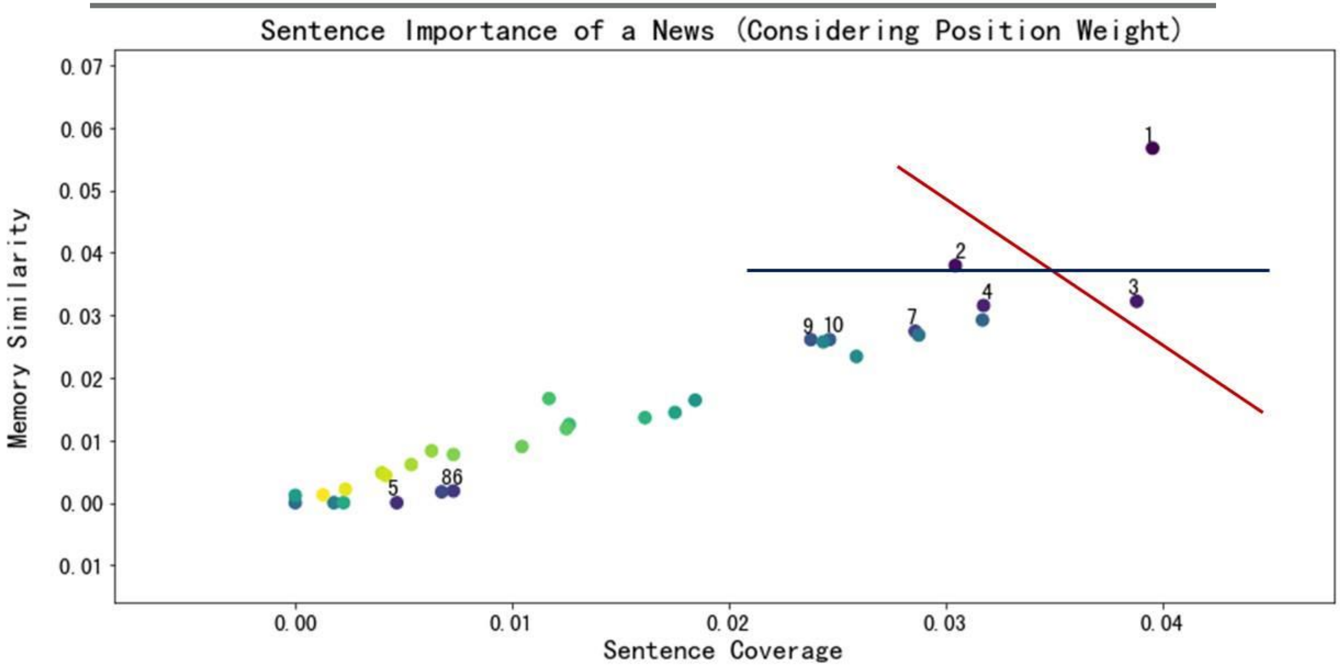
The word embedding is obtained from the open source FastText [62] trained on Wikipedia
The sent. embedding is the average of all available word embedding in the sent.

Error Analysis – Effectiveness of Coverage

懷特（4108）研發上市新藥獲癌症醫界肯定[1]，黃耆經萃取分離及高度純化、研發成功的黃耆多醣注射劑[2]，經證實六成以上患者治療後可緩解癌因性疲憊症[3]，懷特4月營收858萬元，較上月減少21.57%...

Headline :懷特新藥獲肯定，六成患者治療後可緩解癌因性疲憊症。

Newly invented medicine by Phytohealth is praised. Cancer-related fatigue of near 60% patients can be eased after the treatment.

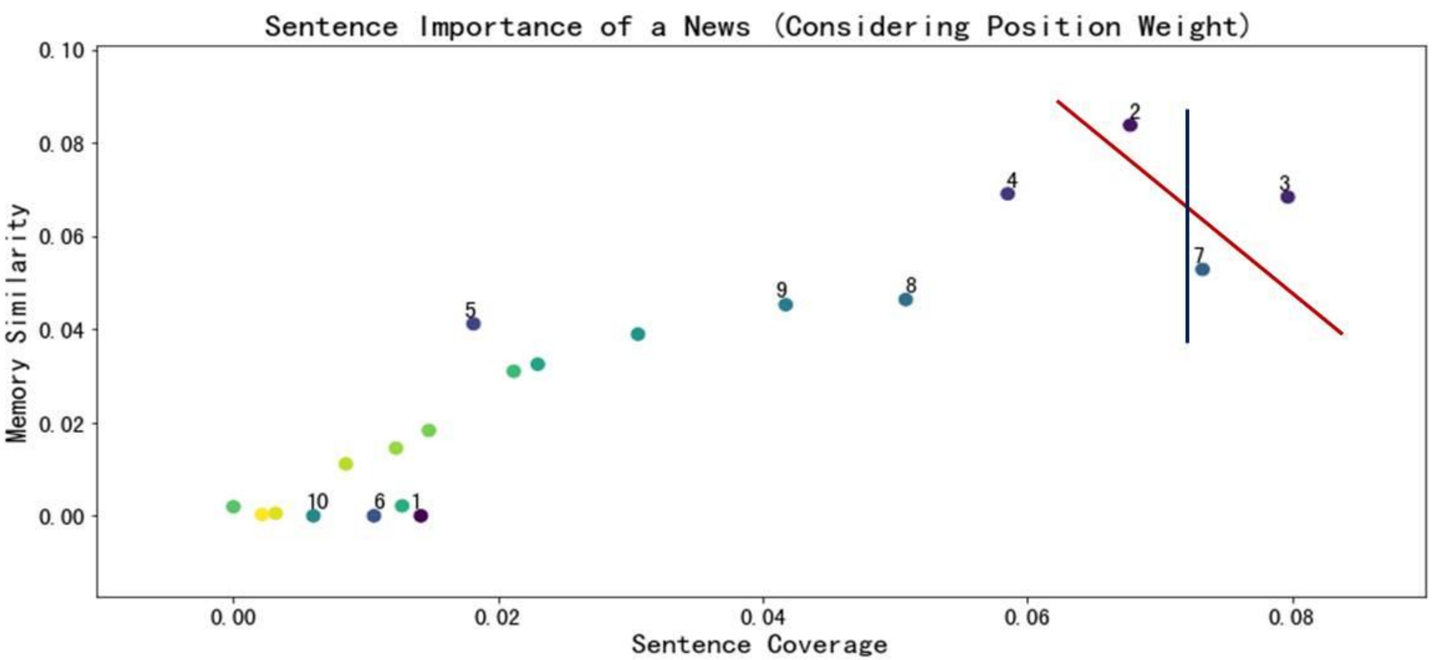


Error Analysis – Effectiveness of Memory Sim.

因應瑪莉亞颱風襲台，台鐵上午9時宣布今（10）日16時前[2]，全線各級列車照常行駛[3]，中午12時會公布16時以後的列車行駛情形，台鐵表示，自發布海上颱風警報起至解除海上颱風警報止，購買上述期間內各級列車車票之旅客[7]，可自乘車日起一年內，持未經使用之車票至各

Headline:台鐵16時前全線列車照常行駛。

Taiwan railway announced the railway system would operate as usual before 4 p.m.

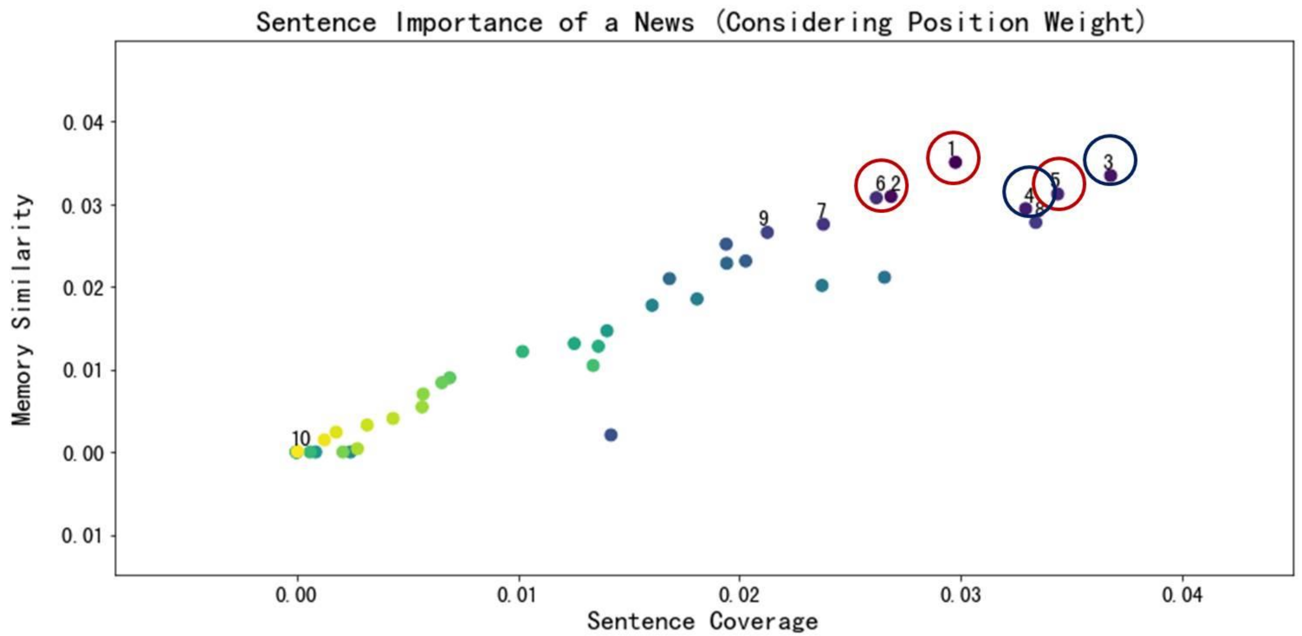


Error Analysis – Revise from Coverage

台視、東森「鐘樓愛人」端午連假仍加緊拍攝中[1]，藝人比莉神秘探班，為兒子周湯豪加油打氣，[3]劇中演員林帥甫也帶著粽子慰勞劇組人員[4]，還爆料周湯豪會畫專屬Q版人物給他[5]，端午節連假「鐘樓愛人」雖馬[6]不停蹄拍攝，但也有不少藝人到場探班…The TV series “Love,

Headline：「鐘樓愛人」端午趕拍，周湯豪手繪Q版人物。

The TV series “Love, Timeless” kept filming during the Dragan boat festival. Nick Chou drew exclusive cutie version characters.



Error Analysis – Revise from Coverage

Examine

Wrong sentences’

Dominant Patterns :

({“周湯豪(Nick Chou)” }, 10), ({“林帥甫 (Lin, Shuai-Fu)” }, 9})

Revise from Patterns

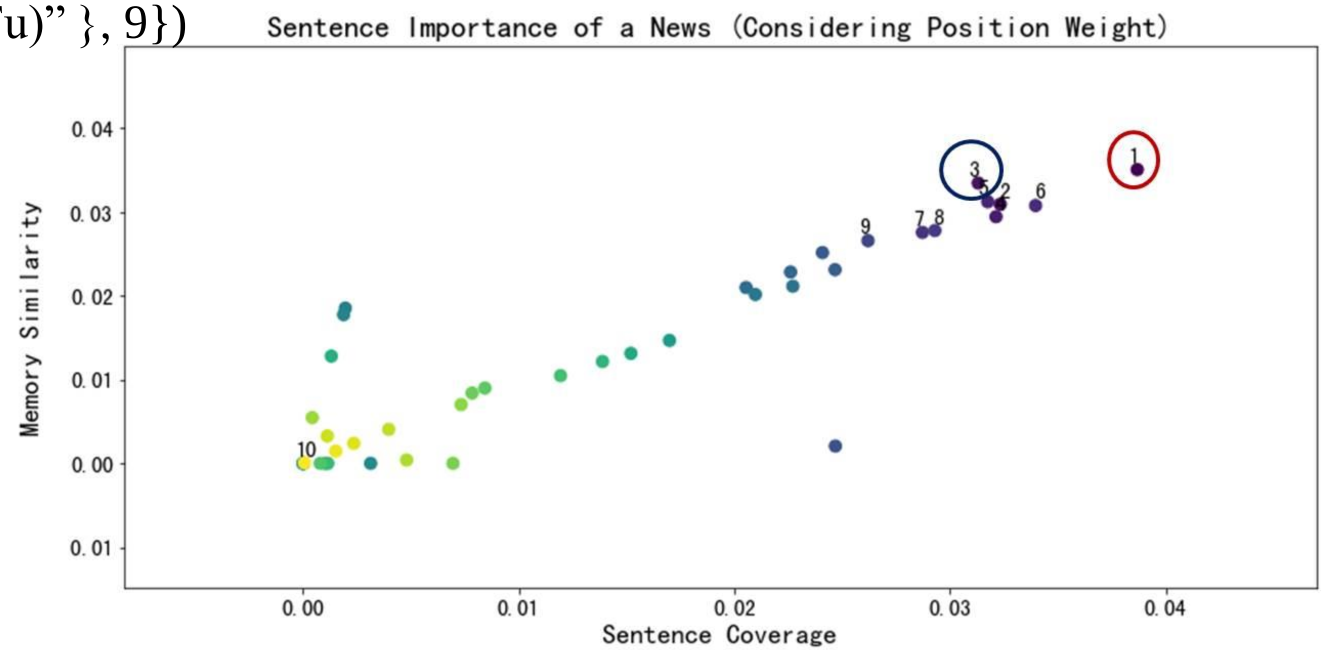
Delete Patterns :

({“周湯豪” }, 10), ({“林帥甫”}, 9)

Retain Patterns : ({“周湯豪”, “林帥甫”},3)

({“劇(TV series)”, “中(In)”, “林帥甫”,} 2)

Repeat same SRCSP PIESim, Scoring process

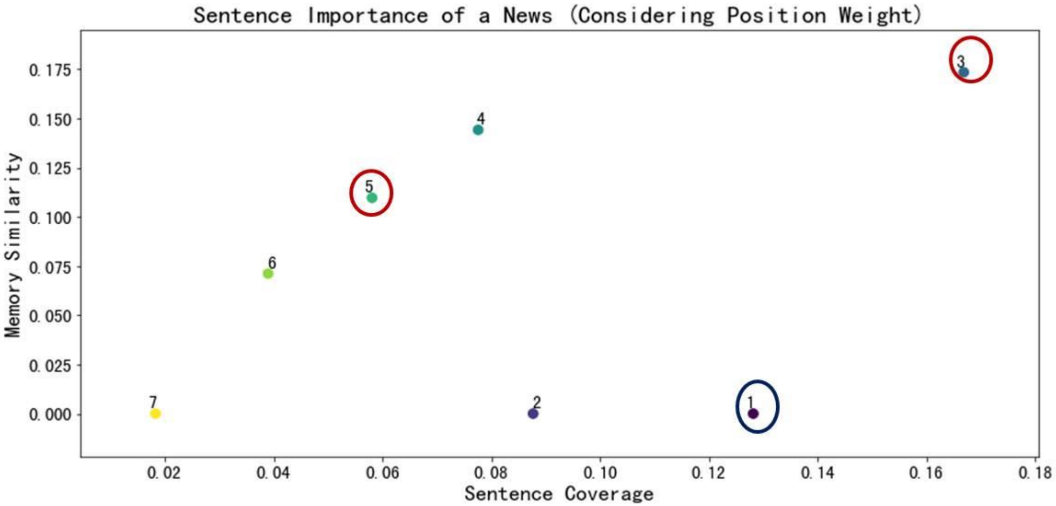


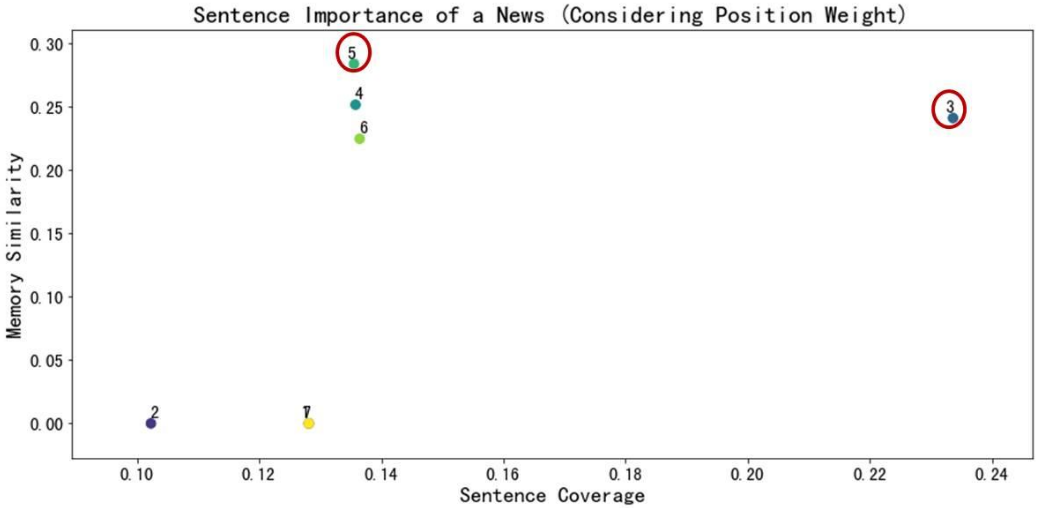
Error Analysis – Revise from Memory Sim.

搶先布局5G商機[1]，遠傳電信董事長徐旭東今(12)日宣布，正式成立「遠傳5G先鋒隊[3]」，攜手工研院、愛立信、國內供應商，打造車聯[5]網創新基地，現場展示兩款自駕車，展現台灣5G應用實力。

Headline :徐旭東宣布成立「5G先鋒隊」主攻車聯網。
Douglas Hsu, announced to establish “Far East vanguard” focusing on Internet of Vehicle.

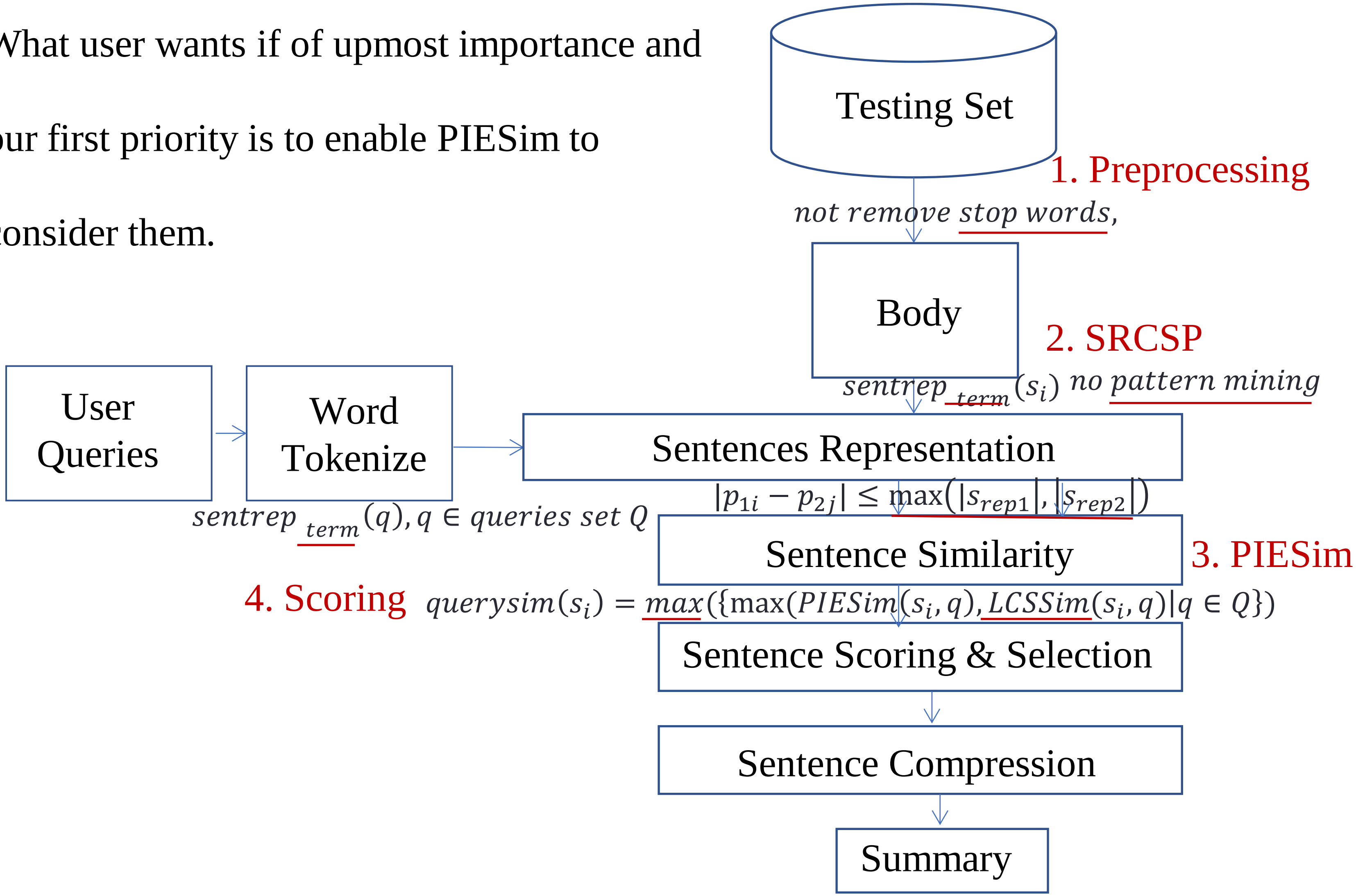
- Limitation : Coverage
- Change 4 least sim. memory
  “組建車聯網 (Build IoV)”
  “5G推動車聯網
  (5G accelerates IoV)”.etc
- Short Text
  Remove position aspects
- Repeat same SRCSP PIESim, Scoring process





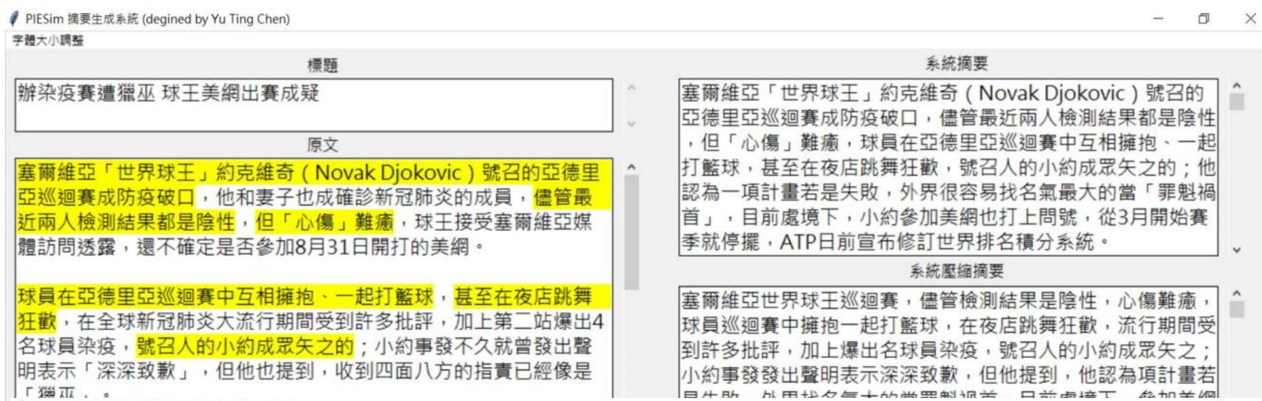
User Interaction – Work Flow

- What user wants if of upmost importance and our first priority is to enable PIESim to consider them.

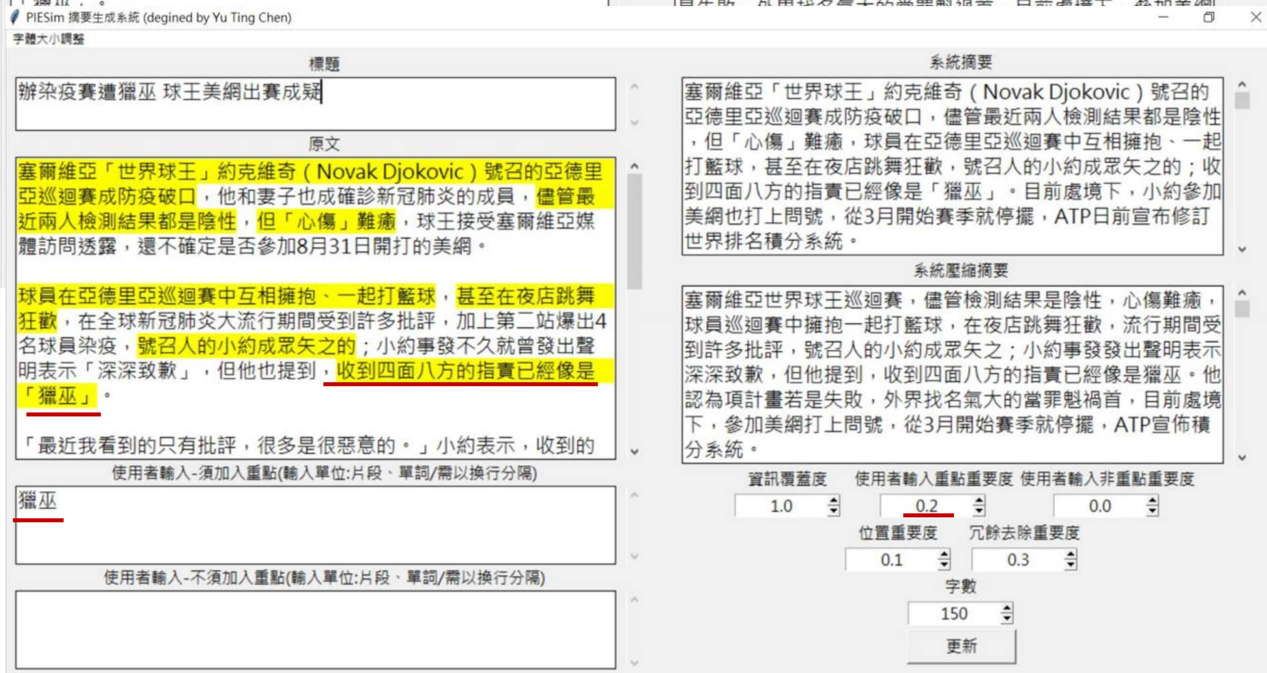


User Interaction – Interface Demo

- Before Query



- After Query



Conclusion & Contribution

Term/ Graph based :
✗ Semantic Info.
✗ Info. outside doc.

Deep Learning:
✗ Explain & Revise

Both:
✗ Conform to how human process text

PIE Sim

Seq. Pat. Mining
+ Edit Distance
= PIESim

+ Non vector based Contextual Info.

+ Conform to how human process text

Result

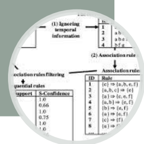
+ Propose novel memory similarity criterion & can consider queries
+ Intuitive to explain & revise
+ Superior performance
+ Innovative mechanism & system

The interaction GUI can be downloaded from https://rb.gy/caqef4. (big since it contains many models & modules). Currently work on **Windows 8/10 64 bytes**. Unzip it and click **the PIESim.exe** in the folder Make your input text **clean will produce a better-quality summary**. Also, since we tested it on UDN, we recommend you to copy news from https://udn.com/news/index, but all formatted articles should work.

Future Work and Prospect

- Sequential Rule Mining

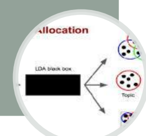
Cause affect relation



CMRules paper

- Topic Modeling or Clustering

Topical Information

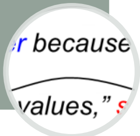


Topic modeling of weather and climate condition paper

- ☐ General and Query-focused
- ☐ Extractive and Abstractive (Chinese)
- ☐ Single document and Multi-document
- ☐ Unsupervised and Supervised

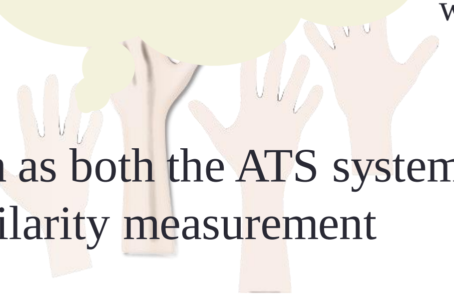
- Linguistic Features ex. co-ref. resolution

Linguistic Techniques



Stanford NLP

bilalqambrani.blogspot



More works using pattern mining and edit dist. In NLP and AI tasks

www.moea.gov.tw

PIESim as both the ATS system and the similarity measurement

Thanks for Your Listening !