KSU Team's System and Experience at the NTCIR-11 RITE-VAL Task

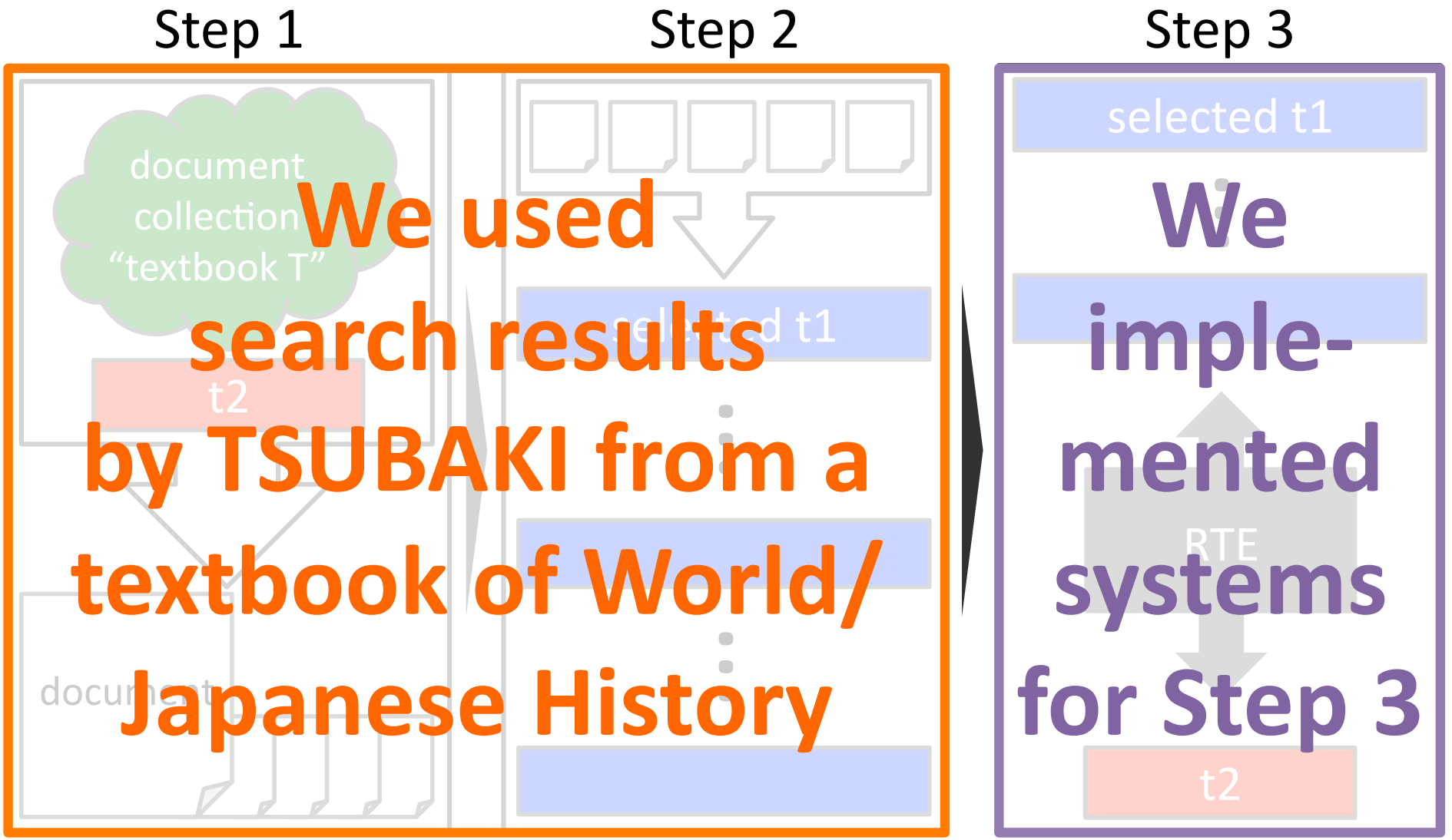
Kyoto Sangyo University, Japan,
Kimura Tasuku, Miyamori Hisashi

NTCIR-11 Conference, Dec 10-12, 2014, Tokyo, Japan

Overview

- Our system for FV
- Our system for SV
- Results of formal and unofficial runs
- Discussion

Fact Validation



System developed for FV

- FV-01
  - System based on character overlap ratio
- FV-02
  - System based on existence of entailment result 'Y'
- FV-03
  - System based on voting of entailment results

Features used by our systems for FV

- Overlap ratio
- Mismatch of expressions
- Strings decomposed into three parts

Overlap ratio

- Overlap ratio of entities

$$overlap(E;t_1,t_2)=\sum_{x\in E}\min(fr(x,t_1),fr(x,t_2))$$

E : a set of entities

$fr(x,s)$: frequencies of x in a given string s

$$overlap_D(t_1, t_2)$$

- Directional overlap ratio of entities

$$overlap_D(E; t_1, t_2) = \frac{overlap(E; t_1, t_2)}{\sum_{x \in E} fr(x, t_2)}$$

e.g.

t1 : 天下統一を目指して信長は、京へ向かった。
t2 : 天下統一を目指した。

$$overlap_D(t_1, t_2) > overlap_D(t_2, t_1)$$

$$overlap_B(t_1, t_2)$$

- Bi-directional overlap ratio of entities

$$overlap_B(E; t_1, t_2) = \frac{2 \times overlap(E; t_1, t_2)}{\sum_{x \in E} fr(x, t_1) + \sum_{x \in E} fr(x, t_2)}$$

e.g.

t1 : 天下統一を目指して信長は、京へ向かった。
t2 : 天下統一を目指した。

$$overlap_B(t_1, t_2) = overlap_B(t_2, t_1)$$

Examples of entities used in FV

- C : set of all **character unigrams** in Japanese
- C^2 : set of all **character bigrams** in Japanese
- K : union of **Kanji and Katakana character** sets

Named entity mismatch

- Mismatch of named entities (NE) in t1 and t2
  - Whether t2 contains named entities not included in t1

True	t1 : 天下統一を目指し、織田信長は京へ向かった。 t2 : 豊臣秀吉は天下統一を目指していた。

False	t1 : 天下統一を目指し、織田信長は京へ向かった。 t2 : 織田信長は天下統一を目指していた。

Number expression mismatch

- Mismatch of number expressions in t1 and t2
  - Whether t2 contains numerical expressions not included in t1

True	t1：織田信長は天下統一を目指していた。 t2：1559年、織田信長は上洛した。

False	t1：織田信長は天下統一を目指し、1559年に上洛。 t2：1559年、織田信長は京へ向かった。

Strings decomposed into three parts

- The longest common prefix h and the longest common suffix t are identified, decomposing the text pair into three parts as follows:

t1 : ファイル交換ソフト開発は違法である。
t2 : ファイル交換ソフトを開発することは違法である。

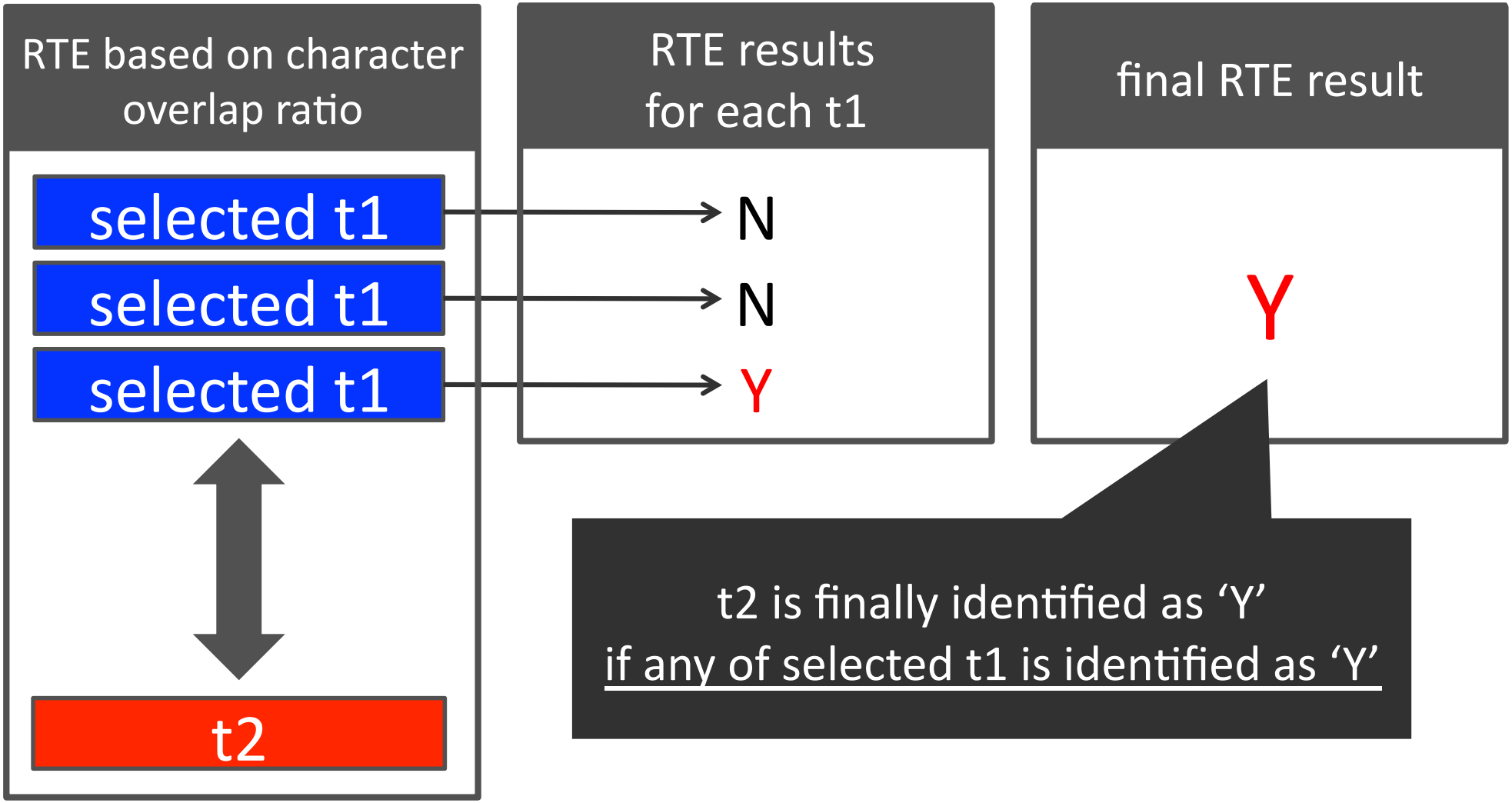
- ht_ratio is defined using h and t as follows:

$$ht\_ratio = \frac{2\left(\left|h\right| + \left|t\right|\right)}{\left|t_1\right| + \left|t_2\right|}$$

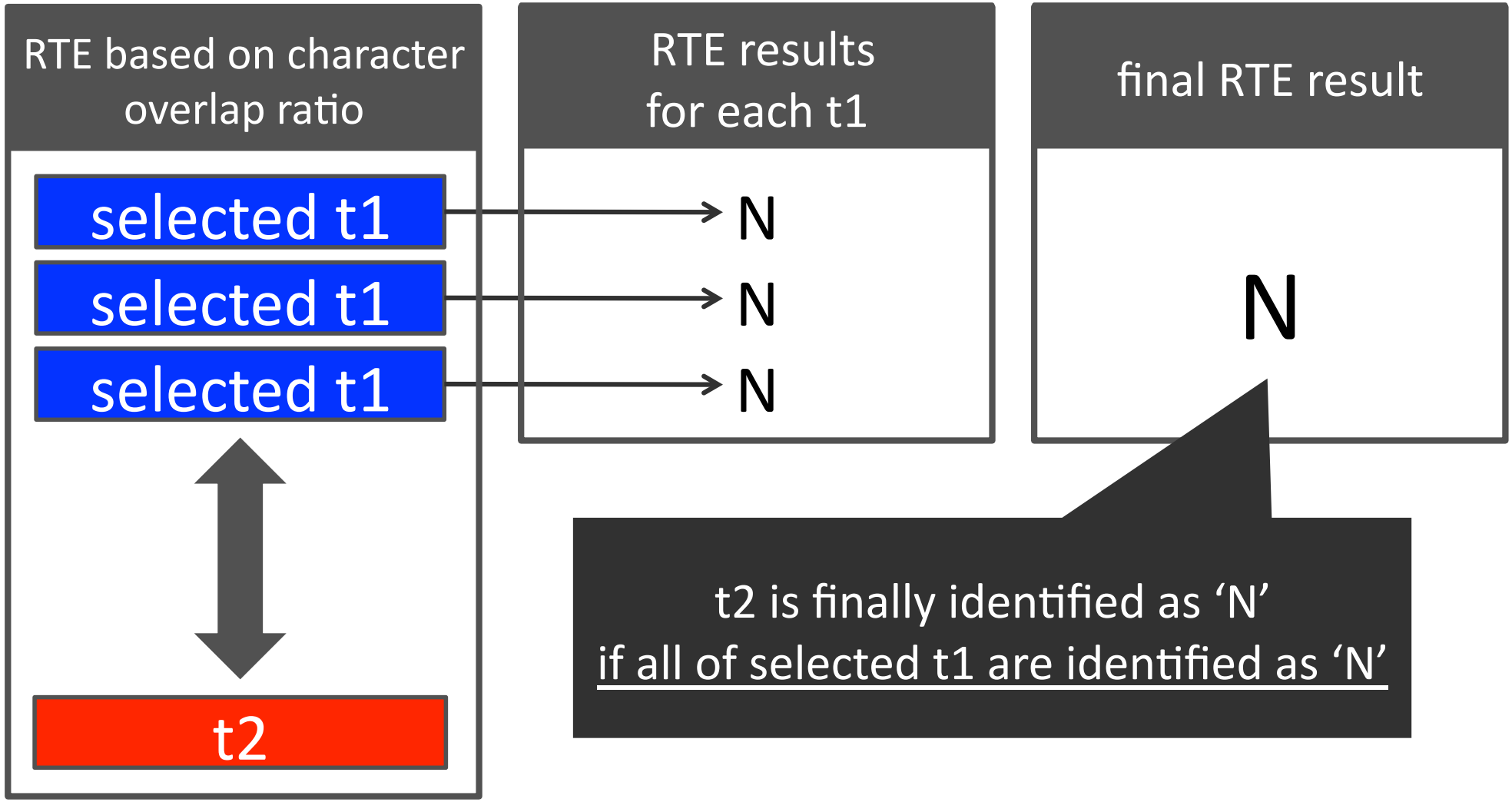
FV-01

- System based on character overlap ratio
- Character overlap ratios are calculated for each of top five sentences of top five documents in search results
- Entailment is identified when **the ratio for any sentence of any document becomes larger than the threshold**

FV-01 : System based on character overlap ratio



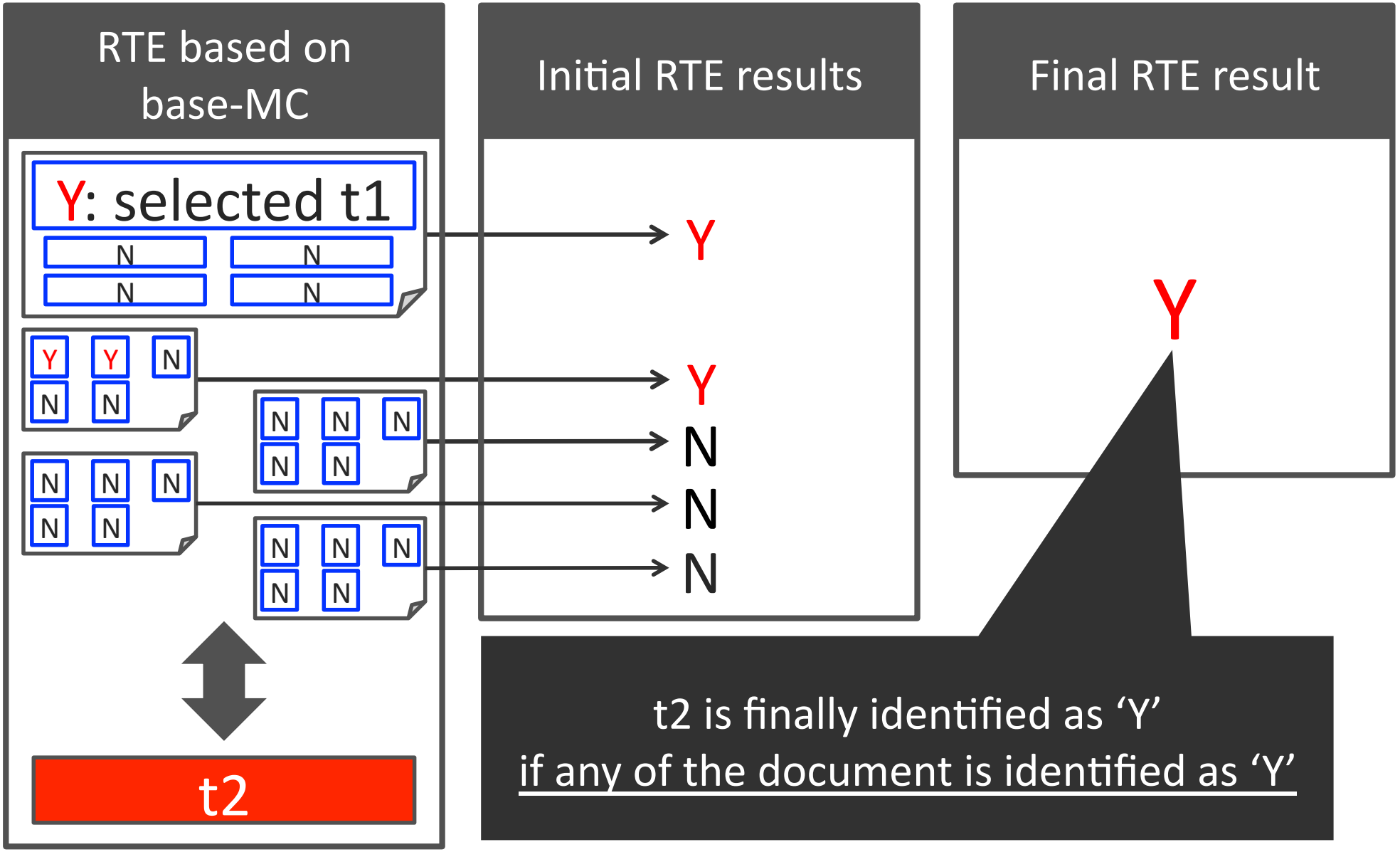
FV-01 : System based on character overlap ratio



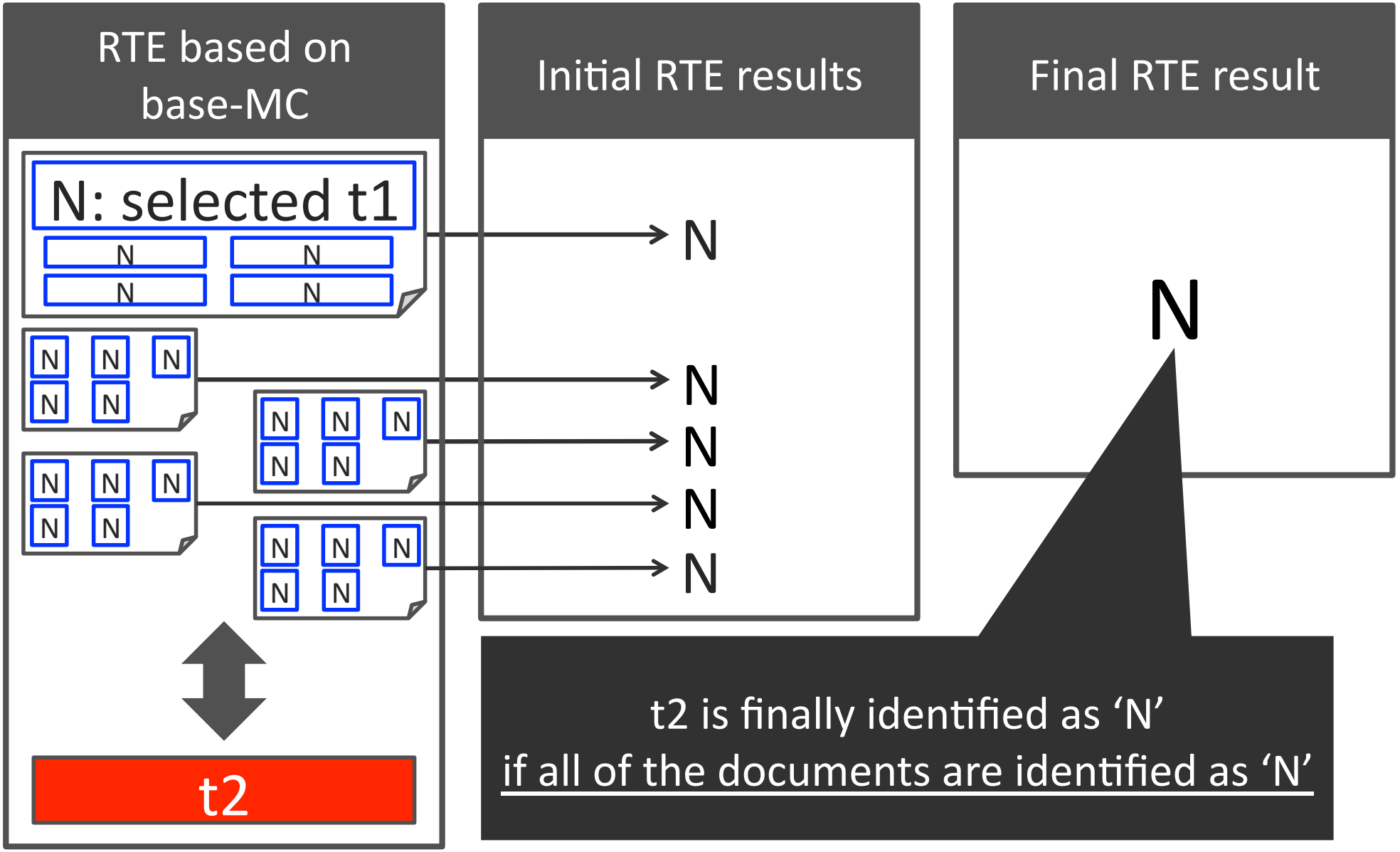
FV-02

- System based on existence of entailment result ‘Y’
- Initial entailment is identified using the base MC for each of top five sentences of top five documents in search results
- Final entailment is identified **if any of the top five documents is identified as ‘Y’**

FV-02 : System based on existence of entailment result ‘Y’

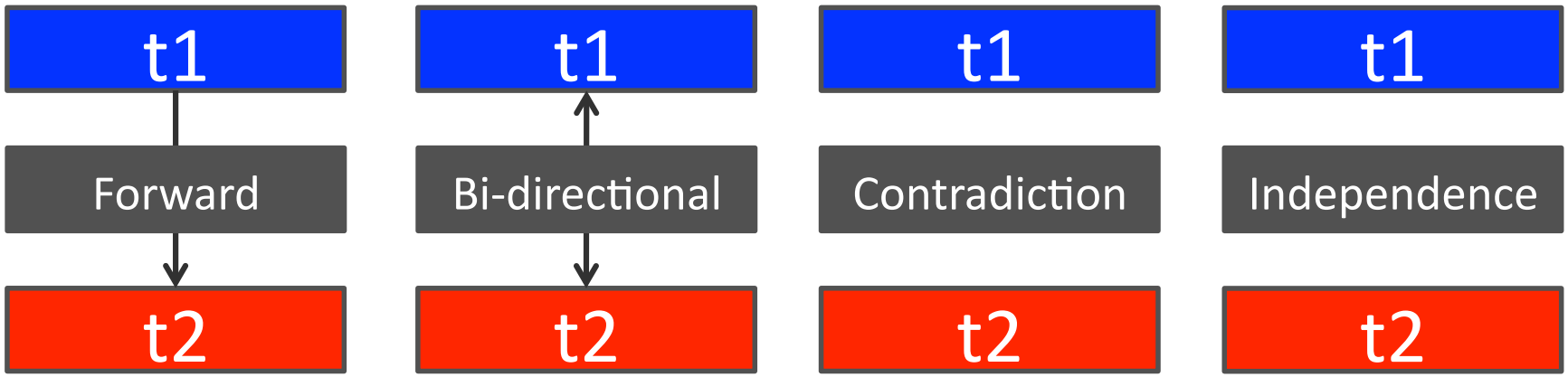


FV-02 : System based on existence of entailment result ‘Y’

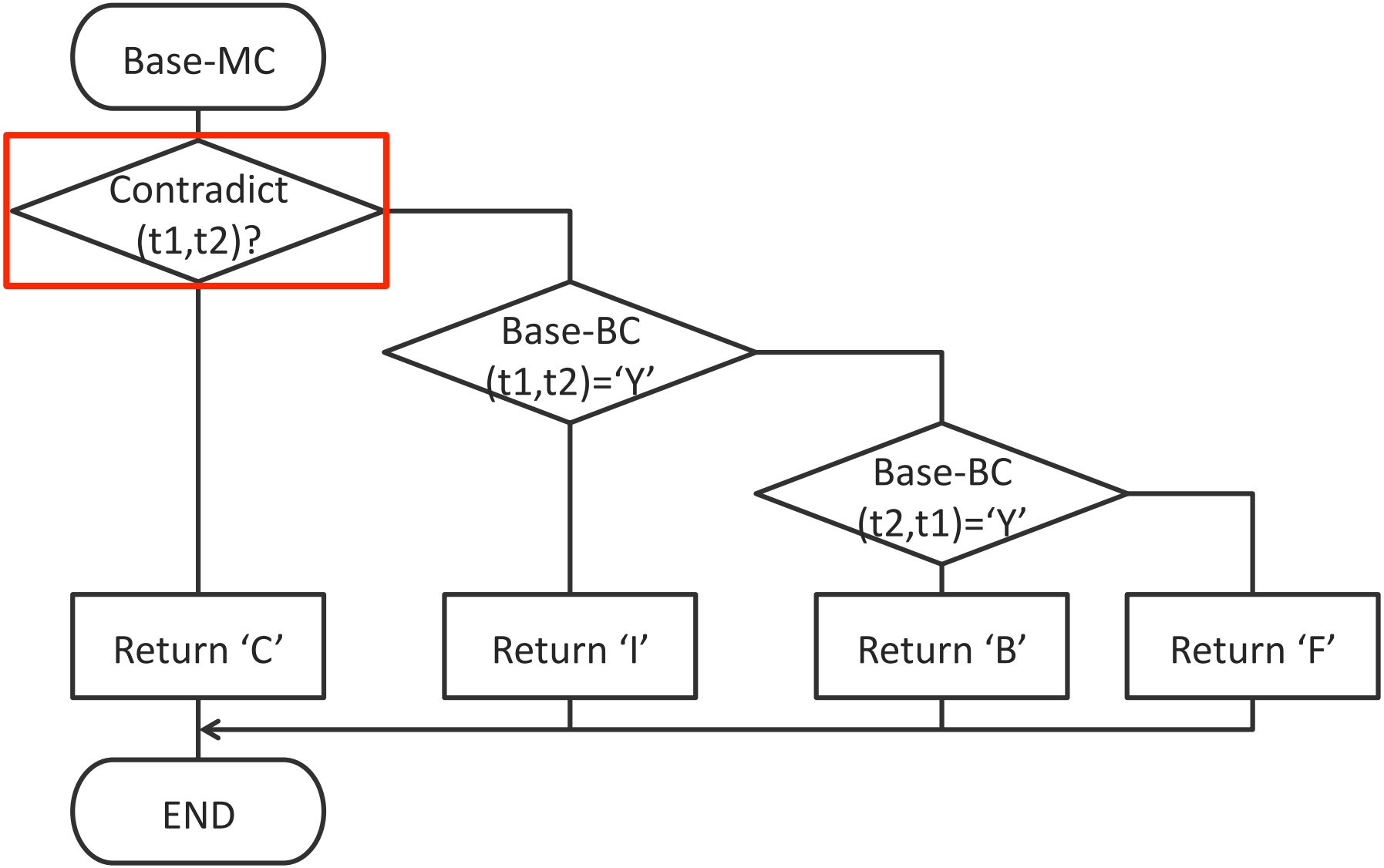


Base-MC

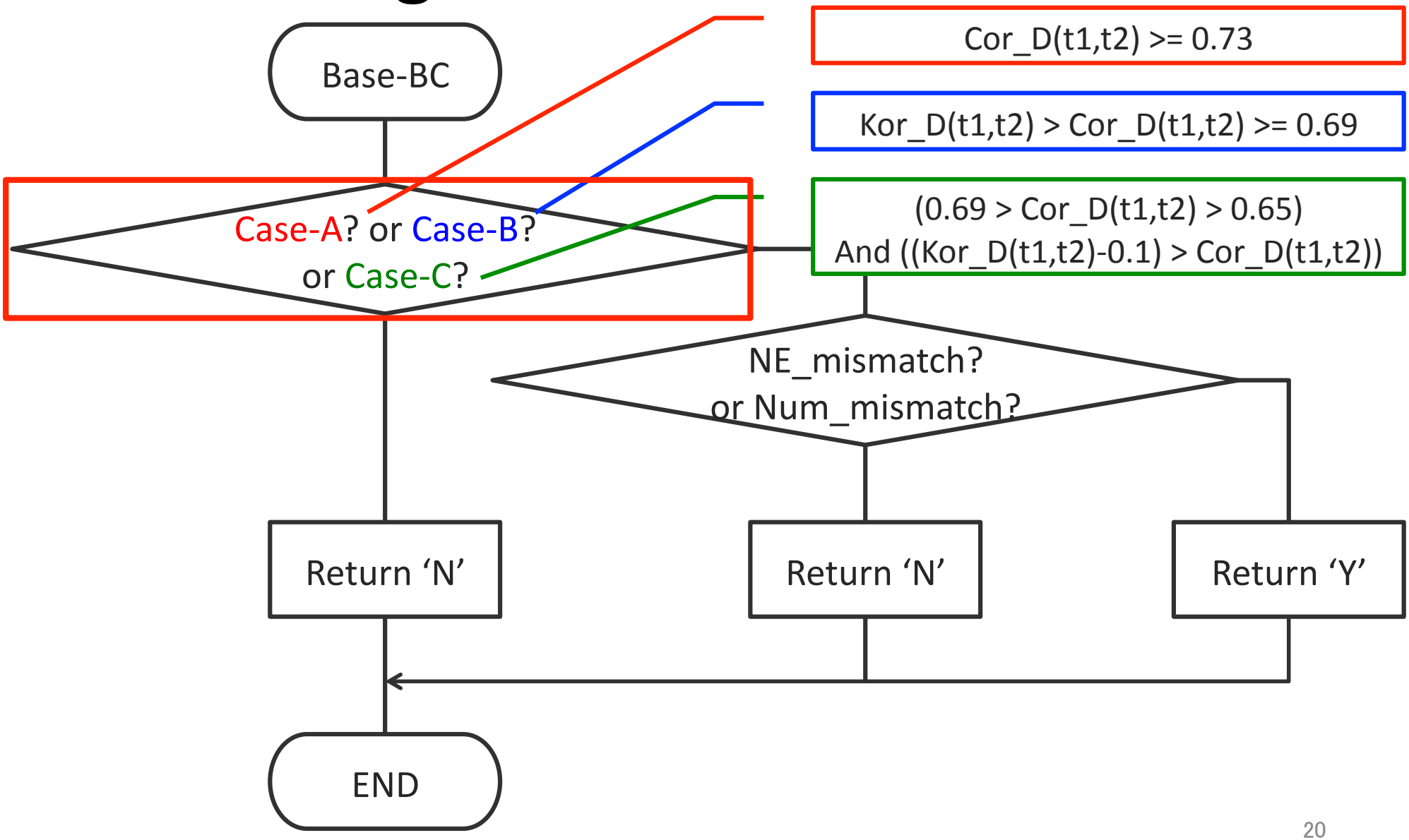
- Base RTE system used for FV-02 and FV-03
- RITE2-SKL-MC-01 was adopted, which gave best performance in MC subtask at RITE2
- MC subtask requires classification into four categories



Algorithm for Base-MC



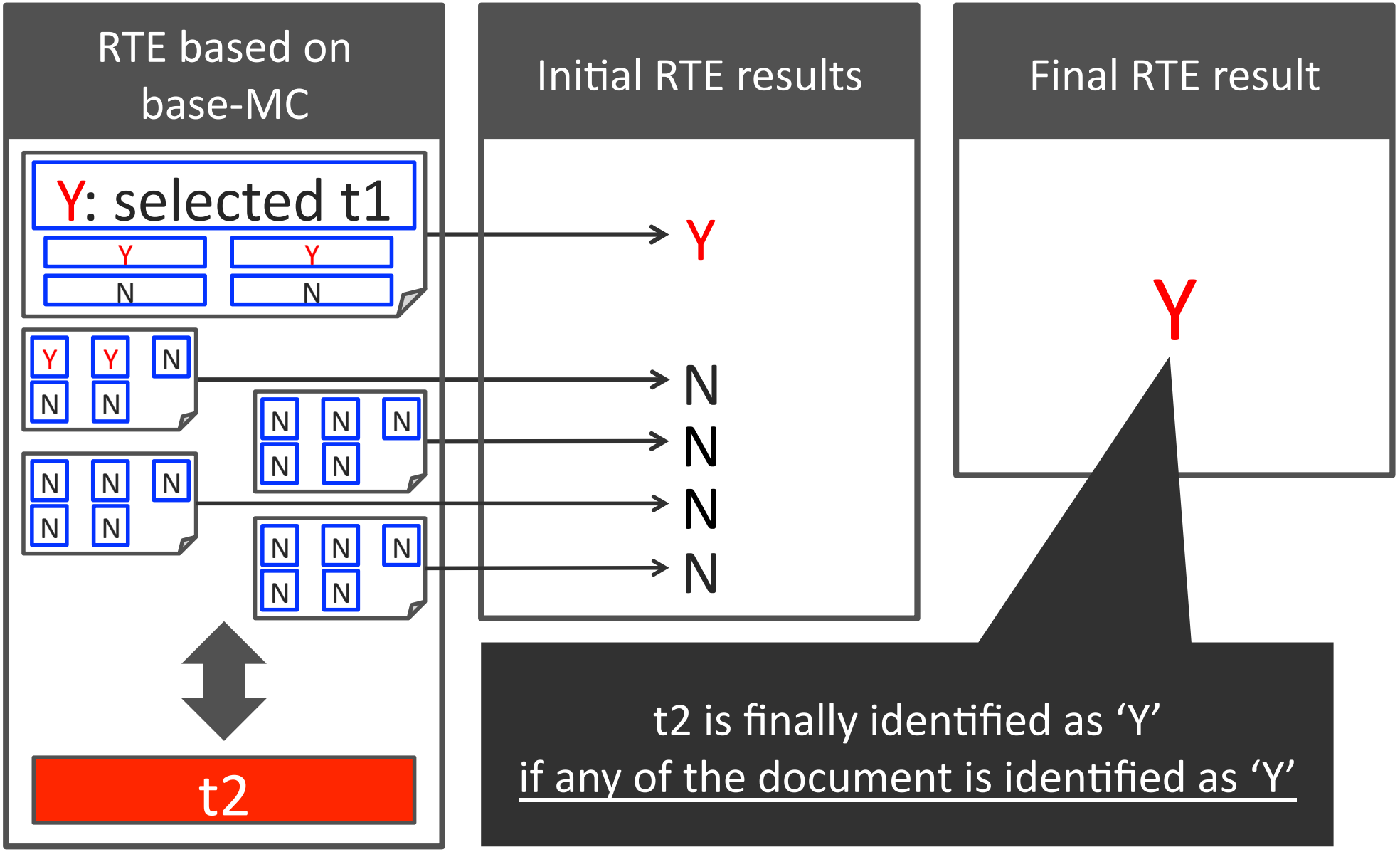
Algorithm for Base-BC



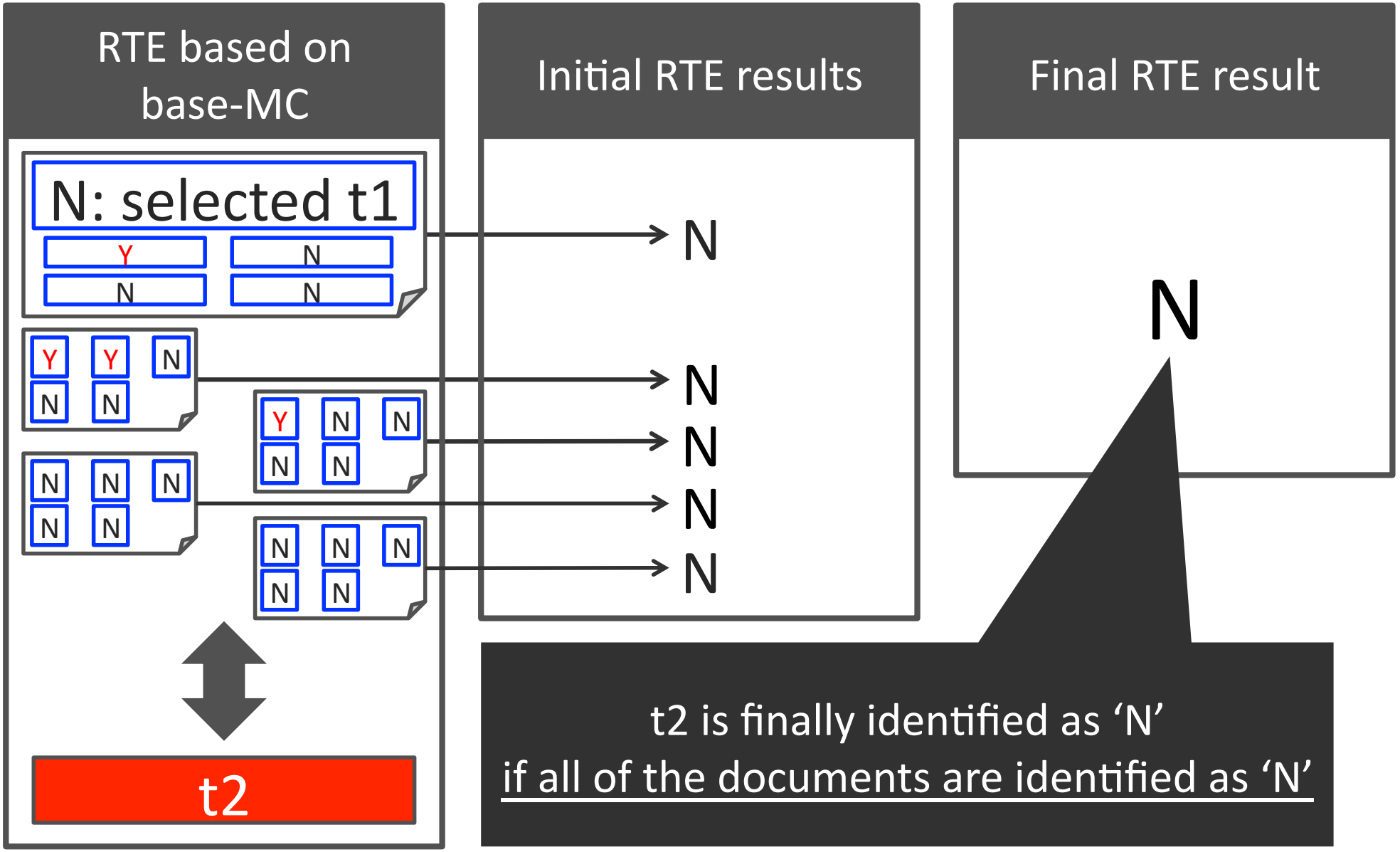
FV-03

- System based on voting of entailment results
- Initial entailment is identified using the base MC by voting the RTE results for each of top five sentences of top five documents
- Final entailment is identified **if any of the top five documents is identified as 'Y'**

FV-03 : System based on voting of entailment results



FV-03 : System based on voting of entailment results



Summary of our systems for FV

- Final RTE result based on initial RTE results

System	Initial RTE	Final RTE	Decision unit
FV-01	character overlap ratio	threshold	sentence
FV-02	base-MC	initial results for sentences	document
FV-03	base-MC	voting of initial results for sentences	document

System Validation

- Task to recognize whether t1 entails t2 for the given text pair
- Requires RTE in various linguistic phenomena related to entailment

System developed for SV

- SV-01
  - System using **SVM** as the classifier
- SV-02
  - System using **Random Forest** as the classifier
- SV-03
  - System using **Bagging** as the classifier

Features used by our systems for SV

- Surface features
- Numerical expression-based features
- Location features
- Named entity features

Surface features

- cos_sim_w
- cos_sim_c
- jaccard_coeff_w
- lcs

cos_sim_w / cos_sim_c / jaccard_coeff_w

- Cosine similarity of **content words**

$$\cos\_sim\_w = \frac{\left|w_1 \cap w_2\right|}{\left|w_1\right|\left|w_2\right|}$$

- Cosine similarity of **characters**

$$\cos\_sim\_c = \frac{\left|c_1 \cap c_2\right|}{\left|c_1\right|\left|c_2\right|}$$

- Jaccard coefficient of **content words**

$$jaccard\_coeff\_w = \frac{\left|w_1 \cap w_2\right|}{\left|w_1 \cup w_2\right|}$$

lcs

- LCS, the longest substrings common to t1 and t2, normalized by the length of t2

t1 : ファイル交換ソフト開発は違法である。
t2 : ファイル交換ソフトを開発することは違法である。

$$lcs = \frac{LCS}{|t_2|}$$

Numerical expression-based features

- numexp_exact
- numexp_n2subset
- numexp_n1subset
- numexp_diff

numexp_exact

- Whether **all the numerical expressions in t2** are exactly included in t1
- Ranges should be the same as those in t2

True	t1：カンファレンスは12/9～12/12に開催される。 t2：12/9～12/12の間、東京へ出張する。

False	t1：12/11にプレゼン発表を行う。 t2：RITE-VALに関する発表は12/10と12/11にある。

numexp_n2subset

- Whether **all the numerical expressions in t2** are partially included in t1

True	t1 : RITE-VALに関する発表は12/10と12/11にある。 t2 : 12/11にプレゼン発表を行う。

False	t1 : 12/11にプレゼン発表を行う。 t2 : RITE-VALに関する発表は12/10と12/11にある。

numexp_n1subset

- Whether **all the numerical expressions in t1** are partially included in t2

True	t1 : 12/11にプレゼン発表を行う。 t2 : RITE-VALに関する発表は12/10と12/11にある。
False	t1 : RITE-VALに関する発表は12/10と12/11にある。 t2 : 12/11にプレゼン発表を行う。

numexp_diff

- Whether **one or more numerical expressions** exist in t2 which do not match with those in t1

True	t1 : 12/9にはEVIAのカンファレンスがあった。 t2 : 12/11にプレゼン発表を行う。

False	t1 : RITE-VALに関する発表は12/10と12/11にある。 t2 : 12/11にプレゼン発表を行う。

Location features

- location
  - Whether **location names** in t2 are also referred to in t1

True	t1：織田信長は京都を目指した。 t2：1559年、信長は京都へ上った。

False	t1：織田信長は上洛を目指していた。 t2：尾張を治めた武将 信長は上洛を目指した。

Named entity features

- ne_n2subset
- ne_diff
- ne_cos_sim

ne_n2subset

- Whether **all named entities** in t2 are partially included in t1

True	t1：天下統一を目指し、織田信長は京へ向かった。 t2：織田信長は天下統一を目指していた。

False	t1：天下統一を目指し、織田信長は京へ向かった。 t2：豊臣秀吉は天下統一を目指していた。

ne_diff

- Whether **a named entity** exist in t2 which is not included in t1

True	t1：天下統一を目指し、織田信長は京へ向かった。 t2：織田信長は豊臣秀吉を家臣とした。

False	t1：天下統一を目指し、織田信長は京へ向かった。 t2：織田信長は天下統一を目指していた。

ne_cos_sim

- Cosine similarity of **named entities**

$$\cos\_sim\_ne = \frac{\left|ne_1 \cap ne_2\right|}{\left|ne_1\right|\left|ne_2\right|}$$

Systems developed for SV

- SV-01
  - System using SVM as the classifier
  - Poly kernel is used
- SV-02
  - System using Random Forest as the classifier
  - Number of trees set to 150
- SV-03
  - System using Bagging as the classifier
  - REPTree is used as a base classifier

Formal runs in FV

System	Macro F1	Accuracy
NUL-JA-FV-03 (1st)	61.47	62.84
NUL-JA-FV-01 (2nd)	59.94	61.67
NUL-JA-FV-05 (3rd)	59.67	61.87
KSU-JA-FV-02	52.78	63.42
KSU-JA-FV-03	52.42	63.23
KSU-JA-FV-01	50.61	50.97

Validity of documents selected as t1

- Degree of coincidence b/w **correct t1** and **t1 selected by each run or by TSUBAKI** is estimated
- Coincidence measured with **all the correct t1 docs** provided by organizers

System	Precision	Recall	F-measure	MAP
FV-01	0.00783	0.00783	0.00006	0.00392
FV-02	0.01364	0.01957	0.00012	0.00740
FV-03	0.01364	0.01957	0.00012	0.00740
TSUBAKI	0.01364	0.02677	0.00014	0.00854

Validity of documents selected as t1

- Degree of coincidence b/w **correct** t1 labeled as 'Y' and **t1 selected by each run or by TSUBAKI** is estimated
- Coincidence measured only with **correct t1 docs** labeled as 'Y' provided by organizers

System	Precision	Recall	F-measure	MAP
FV-01	0.00801	0.00801	0.00006	0.00401
FV-02	0.01395	0.02003	0.00013	0.00757
FV-03	0.01395	0.02003	0.00013	0.00757
TSUBAKI	0.01395	0.02739	0.00014	0.00873

Example of missing documents

t_2		selected t_1	
id	text	id	text
13	国際連合は，パレスティナを分割する案を採択した。	WBS-59	47年，国連はパレスティナを，ユダヤ人国家とアラブ人国家に分割する決議案を採択した。
41	20世紀前半にイランでは，レザー＝ハーンが，カージャール朝を廃してパフレヴィー朝を開いた。	WB-69	大戦中にイギリス・ロシア両軍に占領されていたイランでは，1921年にレザー＝ハーン（レザー＝シャー）がクーデタで政権を握り，イギリスから独立を回復し，25年にはトルコ系のカージャール朝を廃してパフレヴィー朝を開始した。
84	パリ条約でイギリスは北アメリカ植民地の独立を認めた。	WBS-42	イギリスは1783年にパリ条約で北アメリカ植民地の独立を認め，ミシシッピ川以東の広い土地をゆずった。

Remember that the degrees of coincidence in the previous two tables are estimated lower than those with the truly correct documents

Formal runs in SV

System	Macro F1	Accuracy
NUL-JA-SV-04 (1st)	69.59	77.81
NUL-JA-SV-05 (2nd)	68.94	77.96
NUL-JA-SV-01 (3rd)	68.73	77.81
KSU-JA-SV-01	66.96	79.84
KSU-JA-SV-03	65.72	75.78
KSU-JA-SV-02	64.87	76.00

SV-01 : SVM
SV-02 : Random Forest
SV-03 : Bagging

Unoffical runs in SV

- After submitting the results of formal runs, **errors were found in calculating some features** used in SV subtask
- Corrected results are shown below

Runs	System	Macro F1	Accuracy
formal runs (submitted)	KSU-JA-SV-01	66.96	79.84
	KSU-JA-SV-02	64.87	76.00
	KSU-JA-SV-03	65.72	75.78
unofficial runs (corrected)	KSU-JA-SV-01-C	66.01	79.48
	KSU-JA-SV-02-C	63.80	75.56
	KSU-JA-SV-03-C	67.18	76.50

Ablation analysis

- Ablation analysis were carried out to clarify the degree of contribution by each feature for each run

Result of ablation analysis: SVM

Feature	System Description	Macro-F1	Δ
	Baseline	66.01	
Surface features	w/o cos_sim_c	65.57	-0.44
	w/o cos_sim_w	60.34	-5.67
	w/o jc_coef_w	63.18	-2.83
	w/o lcs	64.18	-1.84
Location	w/o location	65.98	-0.03
Named entities	w/o ne_cos_sim	65.91	-0.10
	w/o ne_diff	66.01	0
	w/o ne_n2subset	66.08	0.07
Numerical expressions	w/o numexp_diff	66.08	0.07
	w/o numexp_exact	66.01	0
	w/o numexp_n1subset	66.01	0
	w/o numexp_n2subset	66.01	0

Result of ablation analysis: Random Forest

Feature	System Description	Macro-F1	Δ
	Baseline	63.80	
Surface features	w/o cos_sim_c	63.43	-0.38
	w/o cos_sim_w	61.98	-1.82
	w/o jc_coef_w	63.17	-0.63
	w/o lcs	64.60	0.80
Location	w/o location	64.50	0.70
Named entities	w/o ne_cos_sim	64.51	0.71
	w/o ne_diff	64.41	0.60
	w/o ne_n2subset	63.64	-0.16
Numerical expressions	w/o numexp_diff	63.96	0.16
	w/o numexp_exact	62.95	-0.85
	w/o numexp_n1subset	64.70	0.90
	w/o numexp_n2subset	64.13	0.33

Result of ablation analysis: Bagging

Feature	System Description	Macro-F1	Δ
	Baseline	67.18	
Surface features	w/o cos_sim_c	64.51	-2.67
	w/o cos_sim_w	63.61	-3.57
	w/o jc_coef_w	62.89	-4.29
	w/o lcs	64.40	-2.78
Location	w/o location	66.91	-0.27
Named entities	w/o ne_cos_sim	66.31	-0.88
	w/o ne_diff	67.18	0
	w/o ne_n2subset	67.18	0
Numerical expressions	w/o numexp_diff	67.18	0
	w/o numexp_exact	67.18	0
	w/o numexp_n1subset	67.31	0.13
	w/o numexp_n2subset	67.18	0

Discussion of ablation analysis 1

- **High contribution by surface features** confirmed with any classifiers
- **Low contribution by lcs** observed, however, in Random Forest
- It is presumed that the values of **lcs change in a very wide range** compared to other surface features

Result of ablation analysis: SVM

Feature	System Description	Macro-F1	Δ
	Baseline	66.01	
Surface features	w/o cos_sim_c	65.57	-0.44
	w/o cos_sim_w	60.34	-5.67
	w/o jc_coef_w	63.18	-2.83
	w/o lcs	64.18	-1.84
Location	w/o location	65.98	-0.03
Named entities	w/o ne_cos_sim	65.91	-0.10
	w/o ne_diff	66.01	0
	w/o ne_n2subset	66.08	0.07
Numerical expressions	w/o numexp_diff	66.08	0.07
	w/o numexp_exact	66.01	0
	w/o numexp_n1subset	66.01	0
	w/o numexp_n2subset	66.01	0

Result of ablation analysis: Random Forest

Feature	System Descr	Macro F1	Δ
	Baseline		
Surface features	w/o cos_sim_		
	w/o cos_sim_		
	w/o jc_coef_w		
	w/o lcs	64.60	0.80
Location	w/o location	64.50	0.70
Named entities	w/o ne_cos_sim	64.51	0.71
	w/o ne_diff	64.41	0.60
	w/o ne_n2subset	63.64	-0.16
Numerical expressions	w/o numexp_diff	63.96	0.16
	w/o numexp_exact	62.95	-0.85
	w/o numexp_n1subset	64.70	0.90
	w/o numexp_n2subset	64.13	0.33

Indicates the tendency dissimilar to SVM

Result of ablation analysis: Bagging

Feature	System Descr	Macro F1	Δ
	Baseline		
Surface features	w/o cos_sim_c		
	w/o cos_sim_w		
	w/o jc_coef_w	62.[illegible]	-4.29
	w/o lcs	64.40	-2.78
Location	w/o location	66.91	-0.27
Named entities	w/o ne_cos_sim	66.31	-0.88
	w/o ne_diff	67.18	0
	w/o ne_n2subset	67.18	0
Numerical expressions	w/o numexp_diff	67.18	0
	w/o numexp_exact	67.18	0
	w/o numexp_n1subset	67.31	0.13
	w/o numexp_n2subset	67.18	0

Indicates the tendency similar to SVM

Discussion of ablation analysis 2

- Only slight differences were observed in macro F1 when **removing either numerical expression-based features, location features or NE features with SVM and Bagging**
- Some of the macro-F1 and accuracy were decreased as much when **removing each of those features** as when removing surface features with Random Forest
- Some of the **numerical expression-based features, and named entity features** bear an inverse relation, where one feature becomes 'true' when the other one is 'false'
  - for example, a relation b/w numexp diff and numexp n2subset
- Therefore, it was found that removing one of those features **didn't help decreasing the macro-F1 or accuracy and rather increased them**

Discussion of ablation analysis 3

- Low contributions of **numerical expression-based features, location features and NE features** compared to that of surface features
- This is because the **rates of document pairs including missing values in these features** were high in the test data:
  - 28% in numerical expression based features,
  - 40% in location features, and
  - 72% in named entity features
- Actually, numerical expression-based features contribute to the **classification strongly in SVM and contribute supplementarily in Random Forest and in Bagging**, when combining with other features such as location features and NE features

Conclusion

- Described the systems and results by KSU team
- In FV
  - three systems were evaluated, each of which are based on **character overlap ratios, existence of entailment result 'Y', and voting of entailment results**
  - Didn't achieve high recognition results. Lots of work to do including features and classification methods

- In SV
  - three systems were evaluated, each of which uses different classifiers, with **surface features, numerical expressions, location and NE features**
  - Achieved the fourth place in formal run
  - Ablation analyses show that surface features are still influential
  - Appropriate introduction of more semantic features is necessary for further improvement