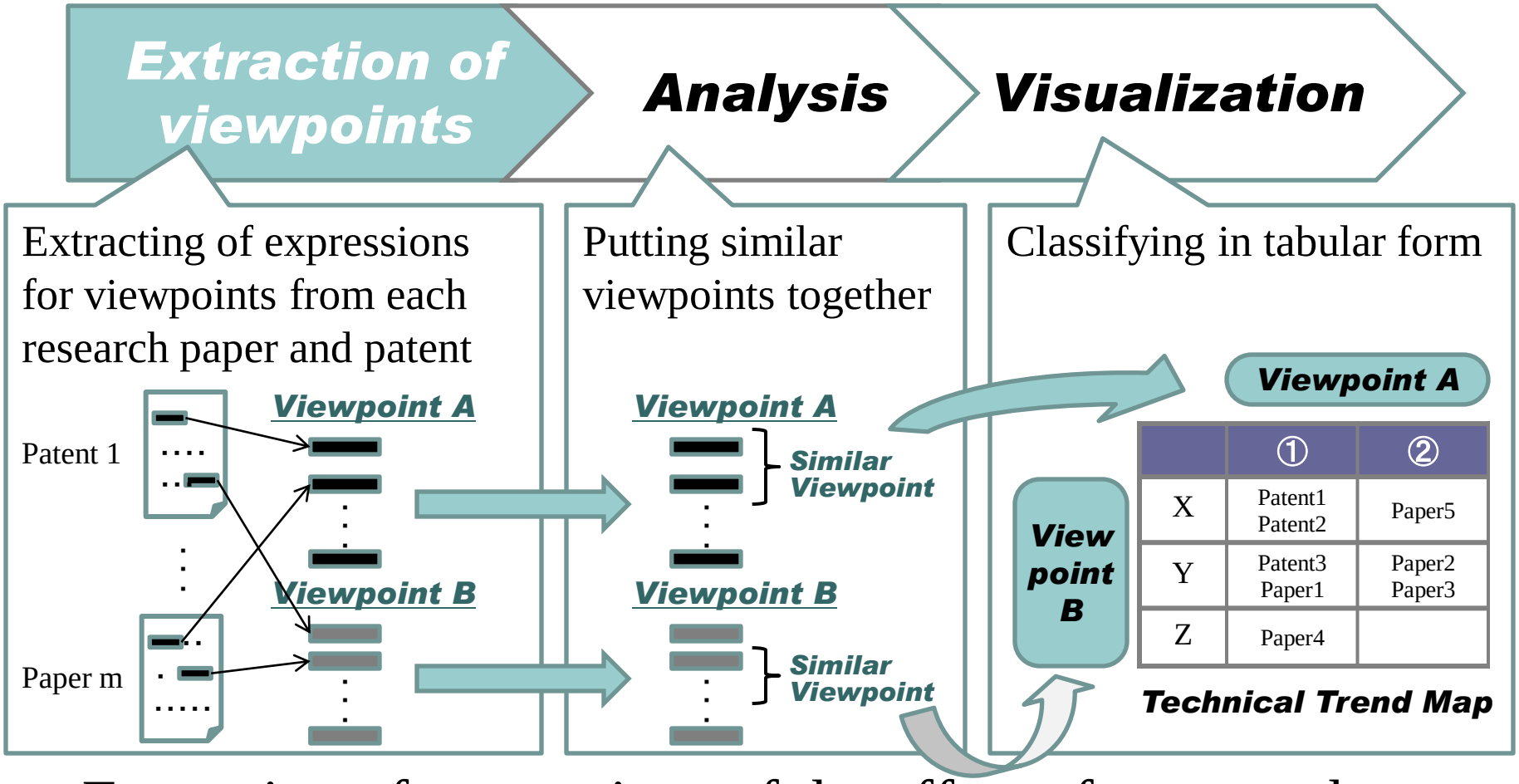
Experiments for NTCIR-8 Technical Trend Map Creation Subtask at Hitachi

Yusuke Sato & Makoto Iwayama

- Creation of Technical Trend Map



	①	②
X	Patent1 Patent2	Paper5
Y	Patent3 Paper1	Paper2 Paper3
Z	Paper4	

- Extraction of expressions of the effect of a research paper and patent as a viewpoint

Purpose

- Difficulty to learn a model for assignment of NTCIR-defined tags
  - Grammatically inconsistent definition of the tags
  - Tendency to assign tags to long phrases
- Definition of a 3-tuple syntactic structure for an effect expression
  - Assigning our independently defined tag set and than converting to NTCIR-defined tag set

Our independently defined tags

<TARGET>圧壊</TARGET><SCALE>強度</SCALE>の<IMPACT>高い</IMPACT>

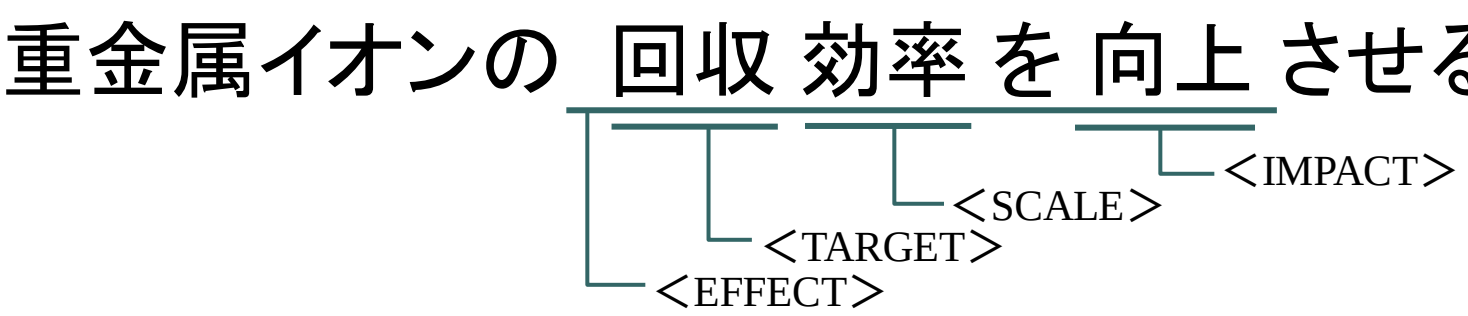
Conversion by several rules

NTCIR-defined tags

<EFFECT><ATTRIBUTE>圧壊強度</ATTRIBUTE>の<VALUE>高い</VALUE></EFFECT>

Our Approach

- Our independently defined tag set



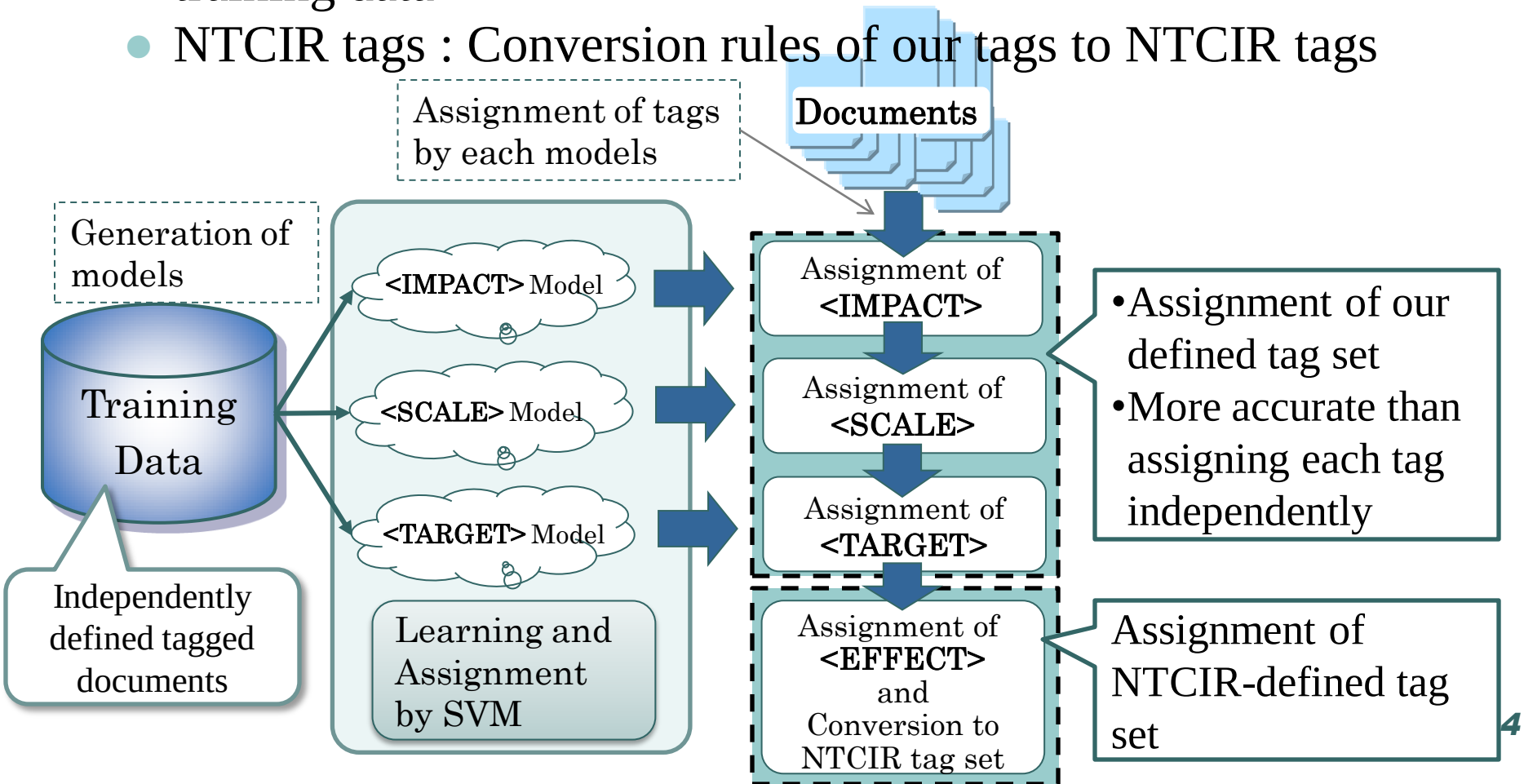
• <EFFECT>	A region including <TARGET>, <SCALE> and <IMPACT>
• <TARGET>	verb or noun which represents an action
• <SCALE>	words such as “速度”, “工程” and so on
• <IMPACT>	words such as “向上”, “低減” and so on

- Difference with NTCIR-defined tags
  1. More consistent grammatical elements
     - <TARGET> : verb or noun, <SCALE> : scale, <IMPACT> : words modifying TARGET and SCALE elements
  2. Division into more common elements or not
     - 回収効率 → specific to some technology fields
     - 回収 → specific, 効率 → common

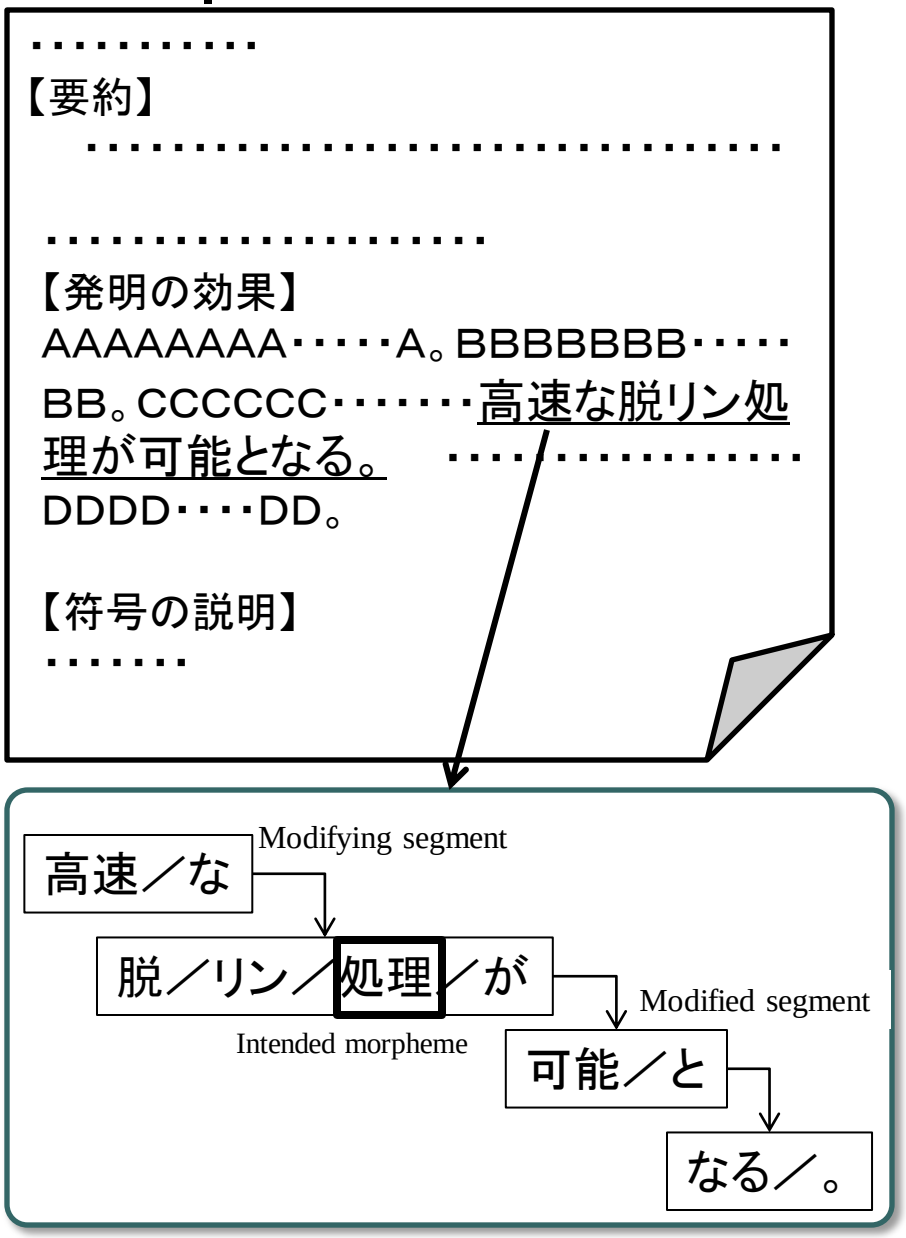
Specific : difficult to assign
Common : easy to assign

The flow of our tag assignment

- Assignment in the order of <IMPACT>, <SCALE> and <TARGET>
- Tag assignment
  - Our tags : Learning by SVM using independently developed training data
  - NTCIR tags : Conversion rules of our tags to NTCIR tags



Features



1. Morphemes by using ChaSen
   脱 ／ リン／**処理**／ が ／ 可能
2. SCALE/IMPACT dictionary
   ⇒ “高速”
3. SCALE/IMPACT-expression prefix/suffix single-kanji
   ⇒ ”高”、”速”
4. Morpheme of head in modifying modified segment
   ⇒ “高速”、”可能”
5. Results of IMPACT/SCALE assignment
   ⇒ ”高速”
6. Information indicating to be effect sentence
   i. End-of-sentence clue-phrase match
      ⇒ ”可能となる”
   ii. Paragraph type
      ⇒ Effect（”効果”）
   iii. Sentence position
      ⇒ 3 /4 = 0.75
   iv. Sentence length
   v. Numeric character ratio within sentence

Assignment of EFFECT tag and Conversion of our tag into NTCIR tag

- <EFFECT> identification

<I>適正</I>な → <S>温度</S>に → <T>制御</T>する → ナビゲーション装置

Along a modification relation

No tags → End of merging

- Merging segments with our defined-tags based on dependency parsing
- Assigning the region to EFFECT tag

<EFFECT><I>適正</I>な<S>温度</S>に<T>制御</T></EFFECT>するナビゲーション装置

Eight rules for converting a combination of out tags to NTCIR tags

- Conversion rules

E.g.) {<I><S>}<T> → <V><A>

- <EFFECT><ATTRIBUTE>適正な温度</ATTRIBUTE>に<VALUE>制御</VALUE></EFFECT>するナビゲーション装置

Independently developed training data

- Training data manually assigned our independently defined tag set

	Data1	Data2	Data3	Data4
Common Data	Abstracts in patent specifications ・Water-purifying technology (C02F 1/28) : 100 ・Learning and classification technology (G06F 17/30) : 98 ・Mixed data A : A61B : 10, B41J : 20, C08L : 10, D01F : 10, E02D : 10, F02D : 10, G06T : 20, H04N : 20			
Extended Data	Mixed data B B : 50 G : 50 H : 50	Mixed data B+ B : 50 G : 200 H : 200	Abstracts in Papers 200	

Data1 : Covering more technology fields
Data2 : Larger volume, but lower reliability for tag assignment
Data3 : For paper
Data4 : Higher reliability, but smaller volume

Experiments

- Features #1 - #3 are commonly used in all runs
- Learning and assignment by SVM (Linear kernel)
  - Giving "+1" if a morpheme is assigned any tag, otherwise "-1"
- No NTCIR-provided training data

#	Type	ID	Training data (Our defined data)	Features		
				#4	#5	#6
i	Patent	HTC_1_1	Data1	✓	✓	✓
ii		HTC_1_2		✓		✓
iii		HTC_2_1	Data2	✓	✓	✓
iv		HTC_2_2		✓		✓
v	Paper	HTC_1	Data3	✓	✓	
vi		HTC_2	Data4		✓	

Results of NTCIR-defined tag set

		Patent				Paper	
		#i	#ii	#iii	#iv	#v	#vi
ATTR.	R	25.1%	24.1%	24.7%	23.7%	14.9%	11.5%
	P	24.1%	23.6%	28.2%	27.3%	16.4%	11.1%
	F	24.6%	23.9%	26.3%	25.4%	15.6%	11.3%
VALUE	R	58.0%	57.2%	52.1%	50.8%	20.7%	23.8%
	P	43.4%	43.2%	46.2%	45.5%	21.0%	20.6%
	F	49.6%	49.2%	49.0%	48.0%	20.9%	22.1%
EFFECT	R	16.4%	15.5%	15.3%	14.5%	5.5%	5.8%
	P	22.3%	21.7%	23.6%	22.8%	11.2%	9.9%
	F	18.9%	18.1%	18.6%	17.7%	7.3%	7.3%
Ave.	R	23.3%	22.7%	21.5%	20.9%	10.0%	10.0%
	P	34.6%	34.4%	38.0%	37.3%	18.8%	16.1%
	F	27.8%	27.4%	27.5%	26.8%	13.1%	12.3%

Results of our independently defined tag set

		Patent (Data1)	Paper (Abstracts in 200 papers)
TARGET	R	45.0%	7.9%
	P	58.7%	19.6%
	F	50.9%	11.3%
SCALE	R	54.3%	19.5%
	P	63.4%	33.8%
	F	58.5%	24.7%
IMPACT	R	64.9%	28.0%
	P	68.4%	38.4%
	F	66.6%	32.4%

Discussion

- NTCIR defined tag set
  - The results of Data1 has slightly higher F-value than those of Data 2
    - Need of higher reliability to tag set rather than a larger volume of data
  - Lower accuracy for papers than patents
    - End-of-sentence clue-phrases in effect sentence are NOT used frequently
- Our independently defined tag set
  - Accuracy of TARGET was low, for which there are relatively few words common to diverse technology fields

Conclusion

- Independent definition of syntactic structure of effect expressions
  - TARGET / SCALE / IMPACT

    <EFFECT><TARGET>建築</TARGET><SCALE>コスト</SCALE>の<VALUE>低減</VALUE></EFFECT>
  - Assignment of our defined tags data by using SVM according to independently developed training
- Conversion of our defined tag set to NTCIR defined tag set by eight rules based on dependency relations
- ATTR. : 24.6%, VALUE : 49.6%, EFFECT : 18.9%
  - “Effect sentence” feature (#6) is very effecttive for patent data
  - Lower accuracy to assign to long phrases