Syntactic Difference Based Approach for NTCIR-9 RITE Task

Team ID: IBM

Yuta Tsuboi, Hiroshi Kanayama, Masaki Ohno
IBM Research – Tokyo

Yuya Unno
Preferred Infrastructure

Submitted results for four Japanese RITE subtasks (BC, MC, EXAM and RITE4QA)

- Performed the best score in **MC** and **EXAM** subtasks

Overview of the approach

- Very difficult task due to the real-world complex dataset

    Impossible to write down hand-crafted rules

        Machine learning approach with various features

- **Similarities** and **differences** between <S> and <T> * are the key features for entailment determination

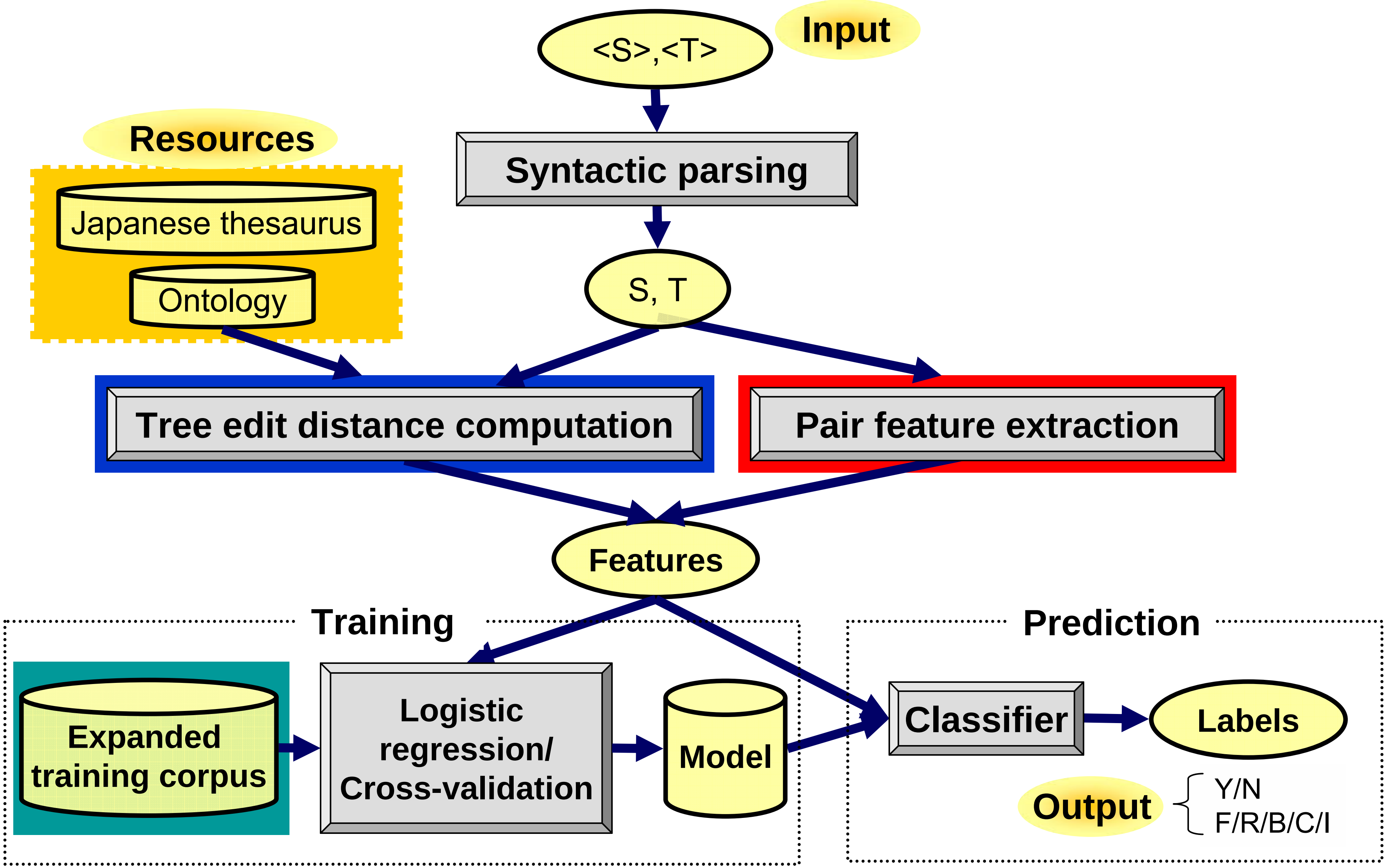
    Matching on the surface is not sufficient

        Calculation of **tree edit distance** between parsed trees

✓ Alignment of two sentences considering syntactic structures
✓ Estimation of similarity (small edit operation costs ⇔ similar pair)
✓ Edit operations (insert, delete and replacement) as features of ML

* <S> and <T> denote an input sentence pair (<t1> and <t2> in the original file)

Techniques and resources for our machine learning approach

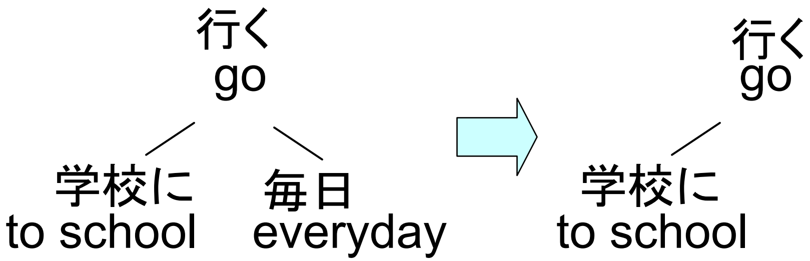


Tree edit distance – Concept

- Hypotheses
  - Two sentences (<S> and <T>) have syntactic similarities and differences
  - A pair of similar sentences has high possibility of entailment
  - Difference parts can be clues for the determination of entailment

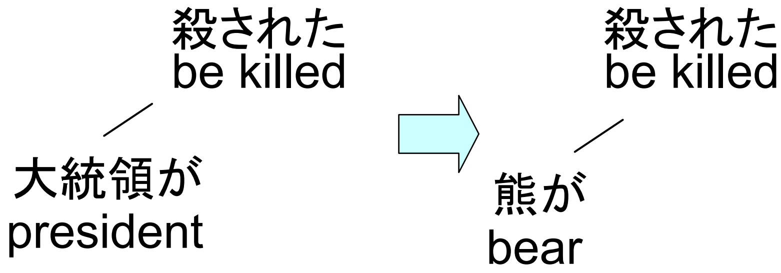
- Solution
  - Parse two sentences
  - Align parse trees by calculating the tree edit distance between them

<S>学校に毎日行く</S>
<T>学校に行く</T>



delete an adverb “everyday”

<S>大統領が殺された</S>
<T>熊が殺された</T>



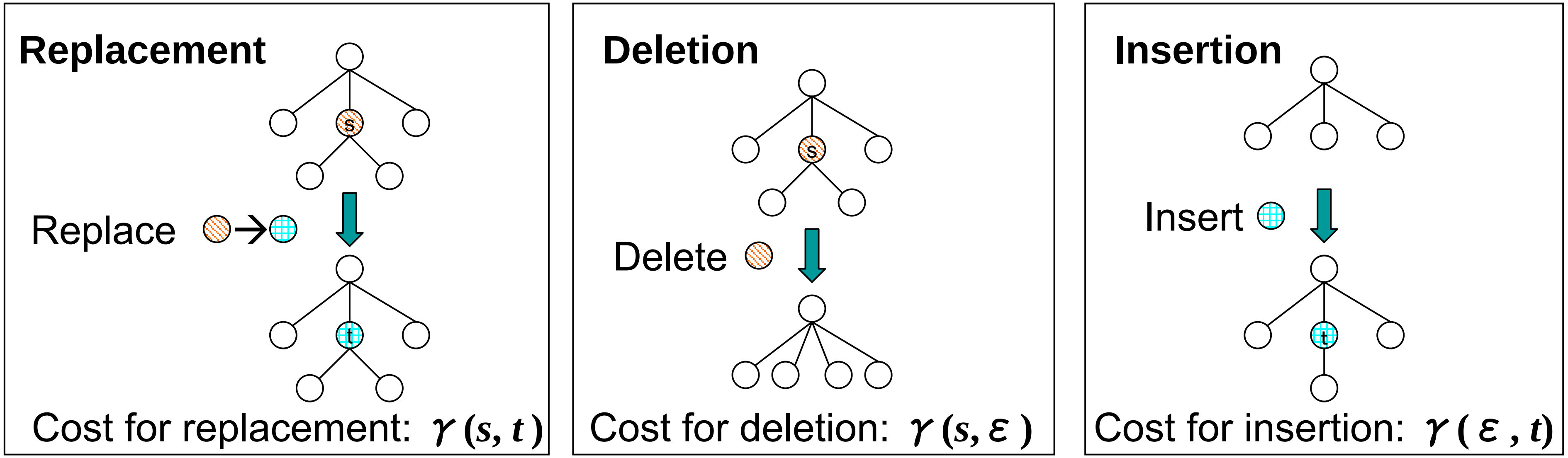
replace “president” into “bear”

Tree Edit Distance – General Implementation

- **Edit distance** δ

$$\delta(\boldsymbol{s},\boldsymbol{t}) = \min_{M} \sum_{(s,t)\in M} \gamma(s,t) + \sum_{s\in D} \gamma(s,\epsilon) + \sum_{t\in I} \gamma(\epsilon,t)$$

- **Edit operations**



- Edit distance computation: O(|**s**|^2|***t***|^2) time and space
- Our code for tree edit distance is available at `https://github.com/unnonouno/tree-edit-distance` (new BSD license)

Cost Functions for Tree Edit Distance

- Insertion / Deletion cost: constant.
  - $\gamma(s, \varepsilon) = \gamma(\varepsilon, t) = 1$
- Replacement cost: a smaller value for a more similar bunsetsu pair
  - Mixed various metrics for similarity:

Cost functions	How to measure the similarity of two bunsetsus
Jaccard distance metrics using word overlap (WO)	✓Overlap ratio between each morpheme set (handling both content and functional words)
Semantic distance metrics using an ontology (Ontology)	✓Inverse of shortest path length of two head content words in the ontology (see next slide)
Semantic distance metrics using thesaurus (BGH)	✓Common depth of two head content words in the thesaurus tree
Heuristic distance metrics (HDM)	✓Similarity value considering parts-of-speech and the thesaurus above

*National Language Research Institute. Bunrui-goi-hyo(revised and enlarged edition), 1996. (In Japanese).

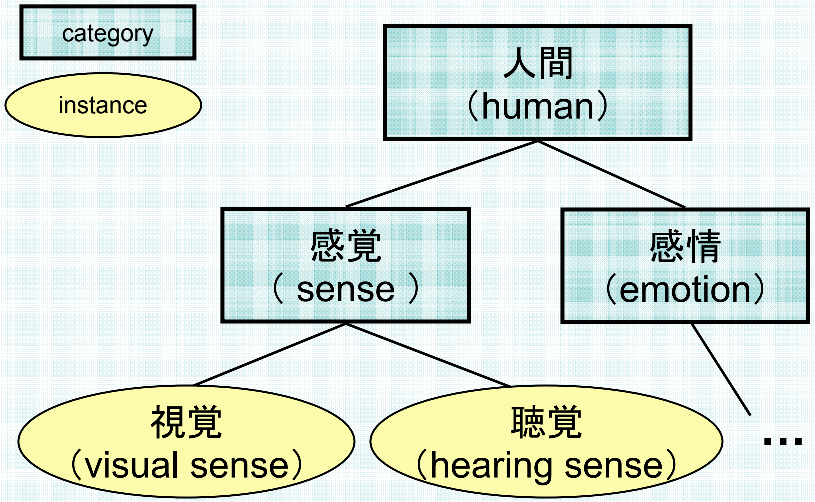
Semantic Similarity and Resources

- We defined two measures for semantic similarity with two complementary resources.

Manually generated resource:

Japanese Thesaurus [分類語彙表]

- High coverage for basic Japanese expression

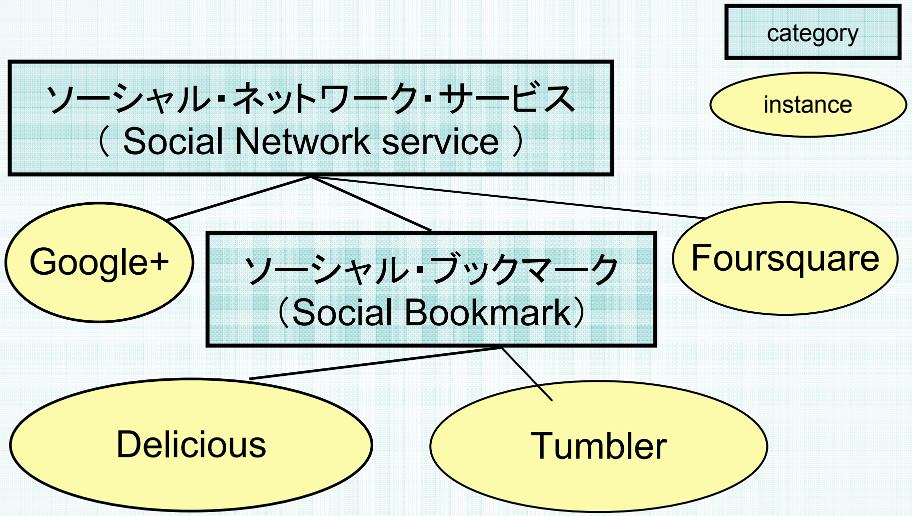


Similarity measure: **Common depth** of two head content words in the tree

Automatically generated resource:

IS-A Ontology generated from Wikipedia *

- Up-to-date knowledge using the latest edition



Similarity measure: **inverse of shortest path length** in the ontology tree

*Y. Shibaki. Constructing large-scale general ontology from wikipedia. Master's thesis, Nagaoka University of Technology, Japan, 2011.

Pair features（1）– Similarity and difference between S and T

- Represent a sentence pair with several features
- Train the logistic regression model using the annotated data

Edit distance and operations (EDO)

0 = same sentence
1 = totally different

Normalized edit distance: $\dfrac{\delta(s, t)}{\max(|s|, |t|)}$

Edit operations:
- insertion / deletion *e.g.* Deletion : "everyday" (word) Deletion : ADVERB (POS)
- replacement *e.g.* Replacement: "president"-"bear" (word pair) Replacement: Noun - Noun (POS pair)

Word overlapping (Word)

Overlap ratio: $\dfrac{m_s \cap m_t}{|m_t|}$

Word pairs: (school, school) (go, school) (everyday, school) (school, go) (go, go) (everyday, go)

All combinations of head content words

Pair features (2) - Ad-hoc strong clues

Sentiment polarity matching

- Applied existing sentiment detector
  - "It is excellent" → positive
  - "I don't like this" → negative
- Sentiment orientation of the sentence pair
  - Same polarity
  - Different polarity
  - Opposite polarity

Strong clue for non-entail

	sentence pairs	features
S	ＰＥＴ（ポジトロン断層撮影）検査は、肺がん、大腸がん、食道がんなど、ほとんどのがんの診療に有効とされている。 'The PET (positron emission tomography) is believed to be effective for the care of most types of cancers such as ...'	$f_{pol} = (+,+)$ $f_{same} = 1$
T	ＰＥＴはがんの診断に役立っている。 'PET helps the care of cancers.'	
S	山田洋次監督は男泣きの場面を作るのがうまい。 'Director Yoji Yamada is good as making scenes of men's weeping.'	$f_{pol} = (+,0)$ $f_{diff} = 1$
T	山田洋次は映画監督です。 'Yoji Yamada is a film director.'	
S	失われた１０年は立ち遅れた反省や経験が生かされ、無駄でなかった。 'The Lost Decade was not wasted because ...'	$f_{pol} = (+,-)$ $f_{diff} = 1$ $f_{opp} = 1$
T	失われた１０年は無駄だった。 'We learned nothing from the Lost Decade.'	

PAS fulfillment test (PAS)

- Convert *S* and *T* to the sets of predicate-argument structures

彼は 大きな 駅へ ゆっくり 行った。
('He went to a big station slowly.')

(P1) 行く (彼, 駅) ('go (he, station)')
(P2) 大きな ⟨駅⟩ ('big ⟨station⟩')
(P3) ゆっくり ⟨行く⟩ ('slowly ⟨go⟩')

- Whether all predicate-argument structures in *T* are covered by those in *S*

Examples of fulfillment

S	スーザン・トレスさんは極めて悪性度の高いがんの一種メラノーマが脳に広がり、脳死になった。 'Ms. Susan Torres became brain dead due to melanoma ...'
T	スーザン・トレスさんは脳死になった。 'Ms. Susan Torres became brain dead.'
S	日本で臓器移植法が施行されて７年以上になる。 'The organ transplantation law have been effective for 7 years in Japan.'
T	日本で臓器移植法は施行された。 'The organ transplantation law became effective in Japan.'

Strong clues for entailment

Pair features (3) – Designed for EXAM subtask

Temporal Matching

- Many sentence pairs in EXAM data includes temporal expressions
- Exploit a feature whether the temporal expressions in S and T have overlap

<S>シュマルカルデン同盟とは、**1531年**に、…</S>
‘Schmalkaldic League is … in **1531**.’

<T>**16世紀**に、ドイツでは、シュマルカルデン同盟…</T>
‘In the **16th century** of Germany, Schmalkaldic League … ’

(year of) 1531
[1531, 1531]

16th century
[1501, 1600]

Temporal Match

Strong clue for entailment

<S>**1997年**まで東南アジアバブルであったが、…</S>
‘Until **1997** there was an economic bubble in South-east Asia, …

<T>**1990年代初め**に起こった日本でのバブル経済の崩壊が、…</T>
‘The collapse of Japanese bubbled economy happened in the early **1990s**, … ’

(year of) 1997
[1997, 1997]

Early 1990s
[1990, 1994]

Temporal Mismatch

Strong clue for non-entail

Expansion of the Training Data

- Convert the training data for the MC subtask as the additional training data for BC subtask, and vice versa.
  - e.g. Forward entailment label (F) between S and T is equal to the true entailment (Y) for S → T and false entailment for T → S.
- Label conversion rule

MC relation	BC relation
$S \xrightarrow{\mathrm{F}} T$	$S \xrightarrow{\mathrm{Y}} T,\ T \xrightarrow{\mathrm{N}} S$
$S \xrightarrow{\mathrm{R}} T$	$S \xrightarrow{\mathrm{N}} T,\ T \xrightarrow{\mathrm{Y}} S$
$S \xrightarrow{\mathrm{B}} T$	$S \xrightarrow{\mathrm{Y}} T,\ T \xrightarrow{\mathrm{Y}} S$
$S \xrightarrow{\mathrm{C}} T$	$S \xrightarrow{\mathrm{N}} T,\ T \xrightarrow{\mathrm{N}} S$
$S \xrightarrow{\mathrm{I}} T$	$S \xrightarrow{\mathrm{N}} T,\ T \xrightarrow{\mathrm{N}} S$

BC relation	MC relation
$S \xrightarrow{\mathrm{Y}} T$	$S \xrightarrow{\mathrm{F,B}} T$
$S \xrightarrow{\mathrm{N}} T$	$S \xrightarrow{\mathrm{R,C,I}} T$

* F,B over arrow denotes either Forward (F) or Bidirectional (B) entailment between S and T (label ambiguities).

- Enhanced data
  - BC+MC' data : 500+880 pairs
  - MC+BC' data : 500+440 pairs (To handle label ambiguities, we train logistic regression by *marginal log*-likelihood maximization)

Results – BC Subtask

Cross validation on training data

Accuracy in formal run

	Cost Function	Features	Training	CV	AC
w/o edit distance	None	Word + Sentiment + PAS + Temporal	BC	52.8	52.0
IBM BC1	HDM	EDO + Word + Sentiment + PAS	BC	54.8	52.2
IBM BC2	HDM + BGH	EDO + Word + Sentiment + PAS	BC	54.0	**52.6**
IBM BC3	HDM + BGH + WO	EDO (POS fine) + Word	BC + MC’	**64.1**	47.0
Oracle	BGH	EDO + Word + Sentiment + PAS	BC	51.8	*56.0*

Best feature set in formal run

- Positive performance gain with the edit distance method
- Very low correlation between CV and AC
  → the results are unpredictable
- BC+MC’ increases CV by 8%～13%, but no effect for AC

Results – MC Subtask

	Cost Function	Features	Training	CV	AC
w/o edit distance	None	Word + Sentiment + PAS + Temporal	MC	33.6	35.9
IBM MC1	HDM	EDO + Word + Sentiment + PAS	MC + BC’	46.8	43.6
IBM MC2	HDM + BGH + WO	EDO + Word	MC	50.2	**50.2**
IBM MC3	HDM + BGH + WO	EDO (POS fine) + Word	MC + BC’	**51.3**	44.5
Oracle	HDM + Ont + WO	EDO + Word + Sentiment	MC	51.1	*51.6*

- Achieved high accuracy for 5-fold classification
- EDO features increased the accuracy by 10%
- Other pair features tend not to work well
- Extended development data (MC+BC’) was not effective

Results – EXAM Subtask

	Cost Function	Features	Training	CV	AC
w/o edit distance	None	Word + Sentiment + PAS + Temporal	EXAM	67.9	69.5
	HDM	EDO + Word + Sentiment + PAS	EXAM	63.7	67.6
IBM EX1	HDM	EDO + Word + Sentiment + PAS+ Temporal	EXAM	**69.1**	**72.2**
IBM EX2	Ontology	EDO + Word + Sentiment + PAS+ Temporal	EXAM	62.5	67.4
IBM EX3	HDM + BGH + WO+ Ontology	EDO (POS fine) + Word + Sentiment + PAS +Temporal	EXAM	61.5	58.4
Oracle	HDM	EDO + Word + Sentiment + PAS+ Temporal	EXAM	68.3	*72.6*

- Relatively small contribution of edit distance
- Temporal increased the accuracy by 5%
- High correlation between CV and AC

Summary

- Achieved good performance in MC and EXAM subtasks
  - Machine learning approach with various features
  - Tree edit operations worked as key features (especially in MC task)
  - Use of thesaurus and ontology – complementary resources

- Performance is unpredictable – the model is still immature
- No special treatment for 5-fold classification in MC task – needed more observation

Backup

Details of MC results – confusion matrix

Correct Label

System Output

	F	R	B	C	I
F	**87**	6	21	30	40
R	7	**89**	20	9	16
B	7	12	**30**	16	6
C	4	0	1	**5**	4
I	5	3	3	5	**14**

Results – RITE4QA Subtask

	Cost Function	Features	Training	AC	Top1	MMR 5
w/o edit distance	None	Word + Sentiment + PAS + Temporal	BC	34.5	5.5	19.8
IBM R4QA1	HDM & Ont & WO	EDO (POS fine) + Word + Sentiment + PAS	BC	33.3	11.3	23.3
IBM R4QA2	HDM & BGH	EDO	BC	31.6	9.1	21.7
IBM R4QA3	HDM & Ont & WO	EDO (POS fine) + Word + Sentiment + PAS +Temporal	BC+MC’	**40.1**	8.7	22.2
Oracle	None	Word + Sentiment + PAS + Tempora	BC+MC’	*63.5*	18.1	*29.0*