NTCIR-11 Conference

NUL System at RITE-VAL tasks

Ai Ishii ,Hiroshi Miyashita,
Mio Kobayashi, Chikara Hoshino
Technology Research & Innovation
Nihon Unisys, Ltd.
2014/12/10

Background

- Text Mining
- Common Sense
- Feature words extraction twitter bot
- UnNatural Language Processing

「空気が読める
コンピュータをつくろう」
プロジェクト

JAPANESE OPEN MIND COMMON SENSE

日本人検定

aimed at high-precision semantic analysis
joined RITE-VAL as bench mark

Agenda

- Introduction
- Shallow approach for RTE
- Search Strategy at FV
- Experimental Results & Discussion
- Future Efforts & Conclusion

Introduction

- Textual entailment recognition (RTE) system for two Japanese subtasks:
  - System Validation (SV)
  - Fact Validation (FV)

- Simple, but robust approach

Overview of our approach

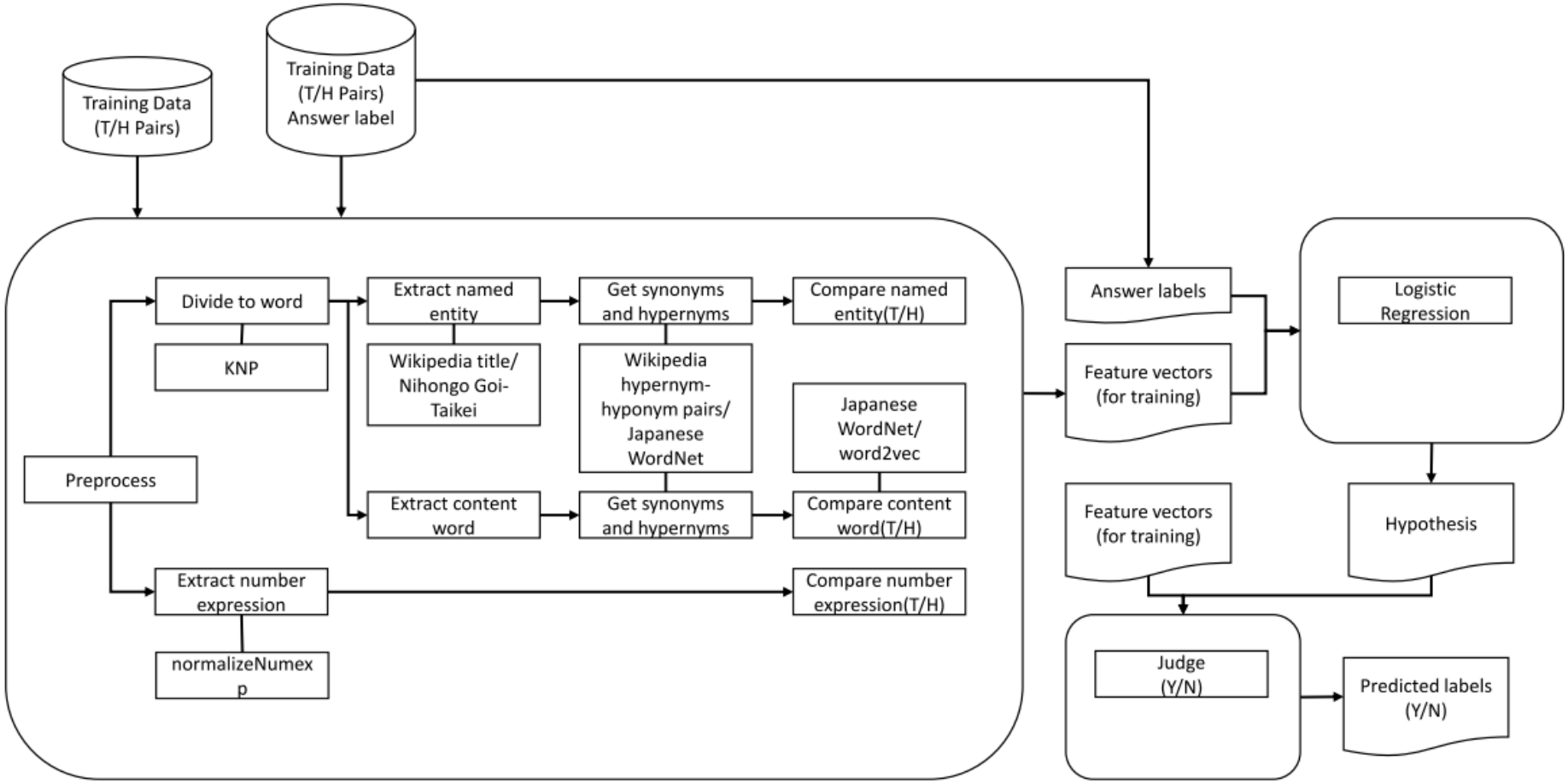
UNISYS

- Shallow apploach for RTE
  - linear classifier mainly based on
    - word overlap feature
    - named-entity feature
  - RITE-2's "A strong shallow system" of team BnO as base system
    - improved named entity extraction
    - transformed some variables
- Apache Solr for FV

Shallow approach for RTE

System Architecture

Feature extraction, learning and classification



Feature Extraction Steps

For a pair of text T and hypothesis H:

1. Chunking
2. Named entity extraction
3. Number expressions extraction
4. Synonym finding
5. Features calculation

Feature Extraction - Chunking

UNISYS

- Chunking
  - divide T and H into word chunks

    using
    - KNP(for SV)
    - Cabocha (mainly for FV)

    word chunk is a independence word and subsequent attached word(KNP's basic clause)

  - identify a word chunk as a Content Word
    - eliminating some stop words
      - "する", "ある", "こと", "もの" : : : etc.

Feature Extraction - Named entity extraction

UNISYS

- Named entity extraction
  - max-length matching from the left
    - for each word chunk in H
    - with concatenating word chunks
  - The knowledge of named entity:
    - Wikipedia titles

      except

      – Nihongo goi taikei ‘s common noun

      – Some exclusion pattern from Wikipedia title list
    - Nihongo goi taikei’s proper nouns

Feature Extraction - Number expression extraction

- Number expression extraction
  - normalizedNumexp
    - extracting number expression
    - converting them into number or date range

    with some hard cording

    example:

    - remove “一つ”(one) from number expression
      - » It is often used, as “one of XX” than as number

Feature Extraction - Synonym finding

UNISYS

- Synonym finding

  For each word in H, find synonym in T

  – The knowledge of synonyms:
    - Wikipedia redirect, Japanese WordNet
    - Nihongo goi taikei
      – for orthographic variation
    - Wikipedia hypernym dictionary
      – by Hyponymy extraction tool
    - Levenshtein Distance
      – for orthographic variation

  For each number expression in H

  – Numerical and temporal entailment recognition
    - number expression of H ⊃ T

Feature Extraction - Features calculation

UNISYS

- Features calculation

  features are defined by:

  – f1: Number expression correspondence
    - f1 = 1 if every number expression in H ⊃ T
    - Otherwise f1 = 0.1

  – f2: Named entity correspondence
    - f2 = 1 if every named entity in H as a synonym in T.
    - Otherwise f1 = 0.1

  – f3: Content word correspondence rate
    - f3 = D_H / L_H
      – L_H: number of words in H
      – D_H: number of words in H that have found their synonyms in T

Feature Extraction- Features calculation

UNISYS

- Features calculation

  features are defined by:

  – Other features

    - f4: Content word first character correspondence rate
    - f5: Word2vec cosign distance
    - f6: Exclusive word
    - f7: Non match content word rate

Learning & Classification

- Input
  - Features and answer labels (Y/N)
- Learning by Logistic Regression
- Classification
  - y = 0 or 1
  - Threshold classifier at 0.5

Search Strategy at FV

Features for FV

Features are defined by:

- f1: Named entity correspondence
  - f1 = 1 if every named entity in H as a synonym in T.
  - Otherwise f1 = 0.1
  - Named entity includes number expression
- f2: Content word correspondence rate
  - $f2 = \log(D_H + 1) / \log(L_H + 1)$
    - L_H: number of words in H
    - D_H: number of words in H that have found their synonyms in T
- f3: Length of H
  - $f3 = \log(L_H + 1)$

Search Strategy

- “Distributed Search” of Solr
  - Search across multiple indexes
    - Wikipedia index
    - textbook index
  - Merge each search results

- Highest-scoring entry as T

Search Strategy

- Unit of search index
  - Search keywords should be near each other regardless of the word order

  - chose paragraphs as unit of search
  - separate by a newline

- Search query
  - weight named entity 5 times

Experimental Results & Discussion

Experimental Results

- SV

id	accuracy	Macro F1	Y-F1	N-F1
NUL-JA-SV-04	77.81	69.59	53.78	85.40

- FV

id	accuracy	Macro F1	Y-F1	N-F1
NUL-JA-FV-03	63.23	61.93	54.89	68.97

Discussion for Effective features

As a result of replication study

- Effective features are follows:
  - f1: Number Expression Correspondence
  - f2: Named Entity Correspondence
  - f3: Content Word Correspondence Rate

Discussion for Search Strategy

UNISYS

- Unit of Search Index
  - Paragraph was very effective
- Search Query
  - Weighting named entity 5 times was not good for test datasets

Future Works

- Deep approach
  - based on syntactic parsing and inference

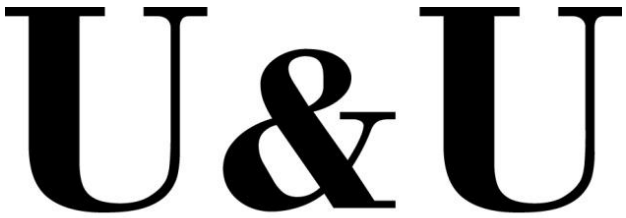
- Keywords specialized in domain
  - Word to be a singleton is different by the domain

    example:
    - "稲作"(rice crop)
    - "貿易"(trade)

Conclusion

- Effective features are follows:
  - Number Expression Correspondence
  - Named Entity Correspondence
  - Content Word Correspondence Rate

- Effective search unit is
  - paragraph by a new line



Users & Unisys

UNISYS

Ex-post Experimental Results of the search method

UNISYS

Effective method as follows:

- Paragraph as unit of search index
- Default weighting of Solr or avoiding length norm
- Sentence as it is for search query

pattern	f1 CC		f2 CC		Macro F1	
	dev	test	dev	test	dev	test
Base line	0.2510	0.2116	0.5212	0.1885	62.59	59.80
paragraph	0.3279	0.2951	0.4611	0.2239	66.46	62.61
paragraph + NE^5	0.3056	0.2711	0.2176	0.1920	65.92	61.93
paragraph - TF	0.3086	0.2871	0.4569	0.2060	65.59	61.64
paragraph - LN	0.3117	0.2945	0.5408	0.2447	65.12	64.28
paragraph + synonym dic	0.3005	0.2771	0.4480	0.2144	65.57	62.49

Base line: page unit index, default weighting, Sentence as it is for search query
CC = correlation coefficient, NE = Named Entity, TF = Term Frequency, LN = Length Norm

Experimental Results of variables

- good variable
  - f1: all named entity has synonym
  - f2: correspondence rate of content word
  - f3: number of words of H

variables	f1	f2	f3	Accuracy		MacroF1	
				dev	test	dev	test
BnO (RITE-2)	1 or 0.1	f1 *log (D_H + 1)	f1 * log (L_H + 1)	64.13	61.87	63.93	61.83
Test-1	1 or 0.1	log (D_H + 1)	log (L_H + 1)	65.06	64.2	63.83	63.17
Test-2	1 or 0.1	log (D_H + 1) / log (L_H + 1)	log (L_H + 1)	65.52	64.59	65.02	64.28

Definition of Solr

UNISYS

- Default Weight

$$w_{t,d} = tf_{t,d} \cdot idf_t^2 \cdot boost_t \cdot norm_d$$

$$= frequency_{t,d}^{\frac{1}{2}} \cdot \left(1 + \log \frac{N}{df_t + 1}\right)^2 \cdot boost_t \cdot norm_d$$

- Length Norm

$$lengthNorm_d = \frac{1}{\sqrt{numTerms_d}}$$

Discussion for Synonym Finding

UNISYS

- Idea: matching of content words loosely
- It is effective, but remaining challenges are a lot

Type	Good Example	Bad Example
Wikipedia Redirect & WordNet JP	UNESCO&国連 増加&上昇 醍醐天皇&延喜	直接&間接
Wikipedia Hypernym Dictionary	ウジェーヌ･ドラクロワ⊂画家 著作権⊂知的財産権	衆議院⊂国会議員 銀行⊂各国 搾取⊂企業
Levenshtein Distance	ユーゴスラビア&ユーゴズラヴィア(dist:3) ゴードン内閣&ゴードン改造内閣(dist:2)	アメリカ&アフリカ(dist:1) ルイ15世&ルイ14世(dist:1)

Comparison with other team

UNISYS

- RITE-2 Data Set

Team	Accuracy	Macro F1	Y-F1	N-F1
Highest of RITE-2	64.51	58.12	41.76	74.48
Our System	65.42	65.12	61.03	69.00

- RITE-VAL Data Set

Team	Accuracy	Macro F1	Y-F1	N-F1
Highest of the other team	57.20	56.57	—	—
Our System	64.59	64.28	60.34	67.38

Exclusion pattern from Wikipedia title list

UNISYS

- Exclusion pattern list by Sekiguchi removed some entry

一覧	^ファイル:	の旗$
における	^日本の	行政区画$
において	県$	^¥d+$
についての	市$	^¥d+世紀$
に関する	区$	^¥d+年
の登場人物	町$	^明治.+年
の歴史	村$	^大正.+年
県立	郡$	^昭和.+年
都立	州$	^平成.+年
道立	出入口$	^¥d+年代$
府立	の大統領$	^¥d+月¥d+日$
市立	の首相$	決議¥d+
区立	の国王$	¥d+条
町立	形電車$	^第¥d+
村立	系電車$	第.+回
曖昧さ回避	駅$	第.+期
^Help:	高等学校$	第.+次
^Category:	中学校$	道.+号.*線
^Template:	小学校$	^オリンピック.+選手団$
^Portal:	幼稚園$	^全国高等学校野球選手権.+大
^プロジェクト:	方法$	

Sekiguchi, Koji, Development of the Natural Language Processing Tools for the Information Retrieval[Project Paper],” JAIST, 2014.