A Machine Learning based Textual Entailment Recognition of JAIST Team for NTCIR9 RITE

Quang Nhat Minh Pham, Le Minh Nguyen, Akira Shimazu
Japan Advanced Institute of Science and Technology

Table and Content

- Introduction
- Related Work
- System Description
- Experimental Results
- Discussion
- Conclusion

Table and Content

Introduction

- RTE is a fundamental task in Natural Language Understanding
- The task is to determine whether the meaning of a hypothesis H can be inferred from the meaning of a text T
- Applications:
  - Question Answering
  - Text summarization
  - Information Extraction
  - Machine Translation Evaluation

Introduction (2)

- NTCIR9-RITE workshop
  - The first RTE shared-task for Japanese, Chinese
  - Four subtasks: Binary class, Multi class, Entrance exam, RITE4QA
- JAIST team participates in three subtasks for Japanese:
  - Binary class (BC)
  - Entrance Exam (Exam)
  - RITE4QA

Introduction (3)

- Overview of the JAIST RTE system
  - Machine-learning-based system
  - Multiple entailment features
  - Make use of Machine Translation for RTE
    - Both translation data and original data are used
    - Determine whether MT can be used to improve the performance of the RTE system

Table and Content

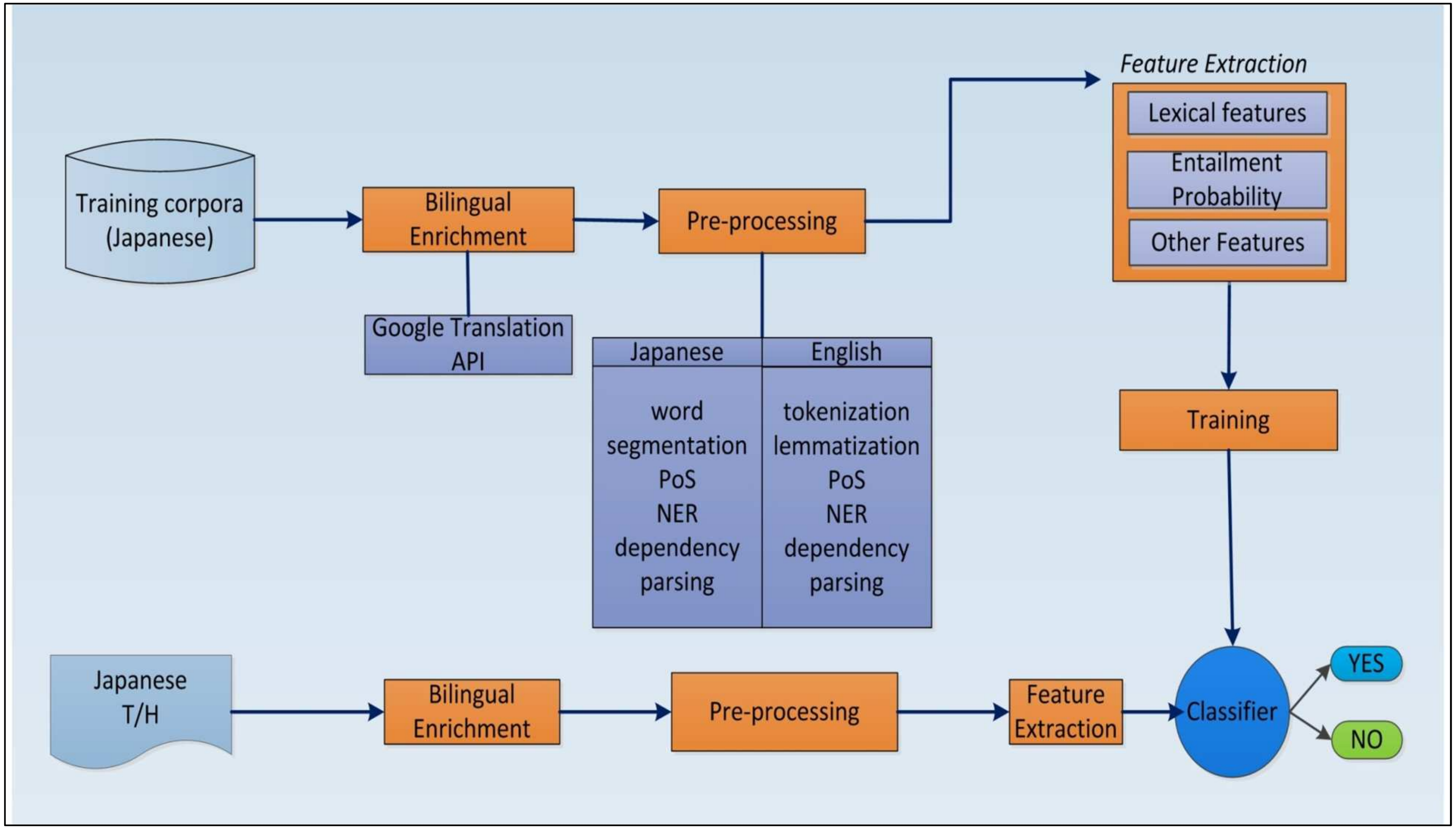
Related Work

- Cross-lingual RTE (Mehadad et al., 2010)
  - Text and Hypothesis are written in different languages
  - A basis solution was proposed
    - A Machine Translation component is added to front-end of an existing RTE system.
- Using bilingual parallel corpora for CLTE (Mehadad et al., 2011)
  - Take advantages of bilingual lexical resources and parallel corpora
  - Phrasal matching

Table and Content

- Introduction
- Related Work
- **System Description**
- Experimental Results
- Discussion
- Conclusion

System Architecture



System Description

- Textual Entailment Recognition as classification problem
- We need:
  - A machine learning algorithm
  - Feature Design
- Bilingual Enrichment
  - Google Translator Toolkit is used to translate Japanese data into English
- Preprocessing
  - Japanese Pairs
    - Cabocha tool
    - Tokenizing, PoS, chunking, named-entity recognition, dependency parsing
  - English Pairs
    - Stanford CoreNLP tool
    - Tokenization, lemmatizaton, PoS, named-entity recognition, dependency parsing

Entailment Classifier

- Machine Learning algorithm
  - Support Vector Machines
- Features:
  - Distance/similarity features
  - Entailment Probability
  - Entailment trigger features
  - …
- Features are extracted from both Japanese pairs and associated English translation pairs

Distance/Similarity Features

- Two representations
  - Surface forms of words in T and H
  - Base forms of words in T and H
- Word overlap
  - For each word in H, find "matching" words in T
  - Compute the number words in H which have matching words
  - Normalize by the length of H
- How to find "matching" words
  - Use English WordNet
    - h_w and t_w have same lemma
    - h_w is synonym of t_w
    - Hypernym, meronym, or member_of distance from t_w to h_w not greater than 3
  - Japanese WordNet
    - h_w is hypernym, meronym, or entailment word (only for verb) of t_w (Japanese WordNet lack synonym relations)

Distance/Similarity Features

- Levenshtein distance
  - Minimum number of edit operations to transform a string to the other
    - Deletion, insertion, substitution
  - Normalization

$$\frac{LevenshteinDist(T,H)}{LevenshteinDist(T,\emptyset) + LevenshteinDist(\emptyset,H)}$$

- BLEU measures
  - Compute n-gram matching between T and H (T is cast as reference translation)
  - Both baseline BLEU precision and modified n-gram precision

Distance/Similarity Features

- Longest common subsequence string (LCS)
  - Compute the length of the longest common subsequence string between T and H
  - Normalize by the length of H
- Other distance/similarity features
  - Jaccard coefficient
  - Mahatan distance
  - Euclidean distance
  - Jaro-Winkler distance
  - Cosine similarity
  - Dice cofficient

Other Features

- Entailment Probability (Glickman et al., 2005)
  - $P(H|T) = \prod_j P(h_j|T)$
  - $P(h_j|T) = max_i \; P(h_j|t_i)$
  - $P(h_j|t_i)$ can be word similarity
    - English: based on Levenshtein distance between two words
    - Japanese: using Nihongo goitaikei
- Dependency-parse based Features
  - Used in paraphrase identification (Wan et al., 2006)
  - Overlap between set of (head-modifier) relations in T and H

$$relation\_overlap = \frac{|relations(H) \cap relations(T)|}{|relations(H)|}$$

Other Features

- Named-Entity mismatch
  - H contains a named-entity that does not occur in T
- Polarity mismatch
  - Only consider root nodes in the dependency parse of T and H
  - Use Polarity Weighted Word List (Takamura et al., 2005)

Table and Content

- Introduction
- Related Work
- System Description
- **Experimental Results**
- Discussion
- Conclusion

Experimental Setting

Table 1: Data statistics

Dataset	Y	N	Total
BC Subtask - Dev set	250	250	500
BC Subtask - Test set	250	250	500
Exam Subtask - Dev set	204	295	499
Exam Subtask - Test set	181	261	442
RITE4QA Test set	106	858	964

- BC subtask and Exam subtask
  - Use development set for training
- RITE4QA subtask
  - Use development set of Exam subtask for training

Official submitted runs

- Submitted runs for BC subtask
  - Run 1 (SVM_bi)
    - Support Vector Machines (libSVM tool)
    - Bilingual features
    - Parameter selection: use the parameter selection tool in the package
  - Run 2 (SVM_mono)
    - Support Vector Machines
    - Use only monolingual features (Japanese)
  - Run 3 (MEM_mono)
    - Maximum Entropy Model
    - Monolingual features
- Submitted runs for Exam and RITE4QA subtask
  - Run 1: Local Lexical Matching (LLM), threshold = 0.65
  - Run 2: SVM_bi
  - Run 3: SVM_mono

Official Results

Table 2: BC Subtask Results

Methods	Accuracy
SVM_bi	0.580 (290/500)
SVM_mono	0.566 (283/500)
MEM_mono	0.552 (276/500)

Table 3: Exam Subtask Results

Methods	Accuracy
LLM	0.622 (275/442)
SVM_bi	0.652 (288/442)
SVM_mono	0.652 (288/442)

Table 4: RITE4QA Subtask Results

Methods	Acc	Top1	MMR5
LLM	0.560	0.180	0.276
SVM_bi	0.676	0.151	0.260
SVM_mono	0.694	0.166	0.273

* Parameters (cost and gamma) affect the accuracy

* Default parameters: SVM_mono: 65.6%; SVM_bi: 69.4% on Exam subtask

Result Analysis

- False-positive pairs
  - System: "Y", Gold label: "N"
  - "N" pairs in which H is highly covered by T in terms of lexical
- False-negative pairs
  - System: "N", Gold label "Y"
  - "Y" pairs which use complicated inference rules (or implicit inference)
- MT component can help to better predict "Y" pairs which have high word overlap

Examples

# ID	Text	Hypothesis	Dataset	Label	SVM_bi
1	石垣島は、冬でもハイビスカスが咲き乱れる楽園だ。 Ishigaki Island is a paradise of bloomed hibiscus even in winter.	石垣島の冬の気温は高い。 Temperature of winter in Ishigaki Island is high.	BC-test	Y	N
9	「イグ・ノーベル賞」（愚かなノーベル賞）の化学賞 に、広瀬幸雄氏（６２）が選ばれた。 Chemistry "Ig Nobel" prize was awarded to Yukio Hirose (62 years old)	ノーベル賞に広瀬幸雄氏が選ばれた。 Nobel" prize was awarded to Yukio Hirose.	BC-test	N	Y
148	主婦や求職中の人も２割いる。 20% of people are housewives and people who are seeking jobs.	２割が、「職場を持たない人」だ。 20% of people are people who do not have workplace.	BC-test	Y	N
28	宝塚歌劇団はチャリティーコンサートを開催した。 Takarazuka Revue Company held a charity concert.	宝塚歌劇団は慈善活動を行った。 Takarazuka Revue Company conducted charity activities.	BC-test	Y	Y

Feature Analysis

- LemmaSim
  - Distance/similarity features computed on base form of each pair T/H
- SurSim
  - Distance/similarity features computed on surface form of each pair T/H
- SynSem
  - Other features: entailment probability, dependency feature, named-entity mismatch, polarity feature

Feature Analysis

Table 6: Feature Analysis

Setting	BC	Exam
SVM_mono + LemmaSim	56.2% (-0.4)	65.1% (-0.5)
SVM_mono + SurSim	56.6% (+0)	64.5% (-1.1)
SVM_mono + SynSem	53.4% (-3.2)	64.0% (-1.6)
SVM_mono + LemmaSim + SurSim	56.8% (+0.2)	64.5% (-1.1)
SVM_mono + LemmaSim + SynSem	56.2% (-0.4)	65.6% (+0)
SVM_mono + SurSim + SynSem	56.0% (-0.6)	66.1% (+0.5)
SVM_mono + All Features	**56.6%**	**65.6%**
SVM_bi + LemmaSim	57.2% (+0.4)	68.1% (-1.3)
SVM_bi + SurSim	57.0% (+0.2)	65.8% (-3.6)
SVM_bi + SynSem	53.4% (-3.4)	65.6% (-3.8)
SVM_bi + LemmaSim + SurSim	58.2% (+1.4)	68.3% (-1.1)
SVM_bi + LemmaSim + SynSem	55.8% (-1.0)	69.2% (-0.2)
SVM_bi + SurSim + SynSem	56.2% (-0.6)	69.9% (+0.5)
SVM_bi + All Features	**56.8%**	**69.4%**

* Default Parameters are used

Ablation Test

Table 7: Ablation Tests

Ablated Resource	BC	Exam
JWordNet	0%	0%
Goi Taikei	0.2%	0.2%
Polarity Words	-0.2%	0.7%
JWordNet + Goi Taikei	0.2%	0.2%
JWordNet + Polarity Words	-0.2%	0.5%
Goi Taikei + Polarity Words	-0.2%	-0.4%
JWordNet + Goi Taikei + Polarity Words	0%	0%

- Conduct ablation test for SVM_mono
- Impact of resources on performance of the system is not significant

Table and Content

Discussion – Entailment phenomena

- True entailment
  - World Knowledge based inference
  - Paraphrasing and entailment words/phrases
  - Hypothesis is a fact extract from text
- False entailment
  - Negation structure
  - Hypothesis discusses an aspect of a topic
  - Ambiguity
  - Wrong inference rules

Examples – True entailment

# ID	Text	Hypothesis
25	歌舞伎は大衆の心をとらえてきた。 Kabuki has captured the heart of the public	歌舞伎は大衆を魅了してきた。 Kabuki has attracted the public.
26	大山のぶ代は『太陽にほえろ！』の脚本家だった。 Oyama Nobuyo is the writer of "Bark at the Sun."	『太陽にほえろ！』の脚本家は女性である。 The writer of "Bark at the Sun" is a woman.
496	６月１０日の「時の記念日」を控え、時計メーカーが、電波を使って正確な時刻に修正する電波時計の新製品を相次いで投入する。 Before the Time Day's June 10th, the watchmaker introduces a series of new products of radio clocks which fix the time exactly using radio waves.	６月１０日は時の記念日だ。 June 10^{th} is the Time Day.

Examples – False entailment

# ID	Text	Hypothesis
17	省エネは、生活レベルを落として原始時代のような生活をしなければならないという思い込みがあるが、そうとも限らないことに気づく必要がある。 There is a belief that in order to conserve energy, we must lower the level of life to primitive ages, but we need to aware that it is not necessary.	省エネは、生活レベルを落として原始時代のような生活をしなければならない。 In order to conserve energy, we must lower the level of life to primitive ages.
188	ベラルーシとポーランドは国境を接し合う隣国同士である。 Belarus and Poland have the same national border.	ポーランドとベラルーシは近隣ではない。 Poland and Belarus are not neighbors.
206	人間の脳は生まれつき、言葉を理解する機能を備えている。 The human brain naturally has the ability to understand language.	人間は動物の中で唯一言語を獲得した。 Human is the only animal can communication by using language
357	日本人の平均所得はイギリスのそれよりかなり上である。 Japanese average income is considerably higher than English people.	日本人はイギリス人より幸福だ。 Japanese people are happier than English people

Conclusion

- Machine Learning based RTE systems
- Features extracted from Japanese pairs and associated English translation pairs
- Does not require deep semantic analysis and extensive engineering efforts
- 58% accuracy on BC subtask
- Major problems:
  - Detecting hard false-entailment examples
    - High lexical overlap
  - Complicated inference rules

Thank you for your listening!