LTI's Textual Entailment Recognizer System at NTCIR-9 RITE

Hideki Shima, Yuanpeng Li, Naoki Orii, Teruko Mitamura
Carnegie Mellon University, USA
{hideki, yuanpeng.li, norii, teruko}@cs.cmu.edu

SUMMARY

Recognizing Textual Entailment is an important and hard basic research that can be applicable to many research fields, e.g. Question Answering, Text Summarization, Information Retrieval and Information Extraction. We, LTI team, developed an adaptable feature-based approach that combines *multiple complementing features* motivated by our analysis on data as well as *linguistic insights*. We evaluated our system in the Japanese tracks of the BC, Entrance Exam and RITE4QA subtasks at NTCIR-9 RITE, which marked relatively good scores among participants.

Task Definition (binary-class)
Input: sentence t_1 and t_2
Output: binary label indicating whether the premise t_1 entails the hypothesis t_2 (meaning that if a human reading t_1 would infer that t_2 is most likely true).

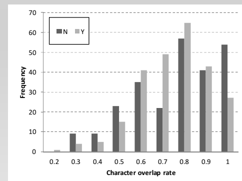
ANALYSIS

Summary of manual analysis on a sample of JA BC dev dataset.

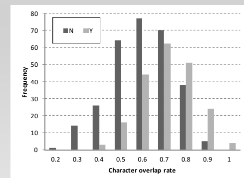
Category	Freq	Example t_1	Example t_2
Lexical Entailment	164	...市町村に移された/moved to a municipality...	...自治体に移された/moved to a autonomous community...
Syntactic Entailment	160	沖縄の基地は、米国の世界戦略と密接に結びついている。/Military bases in Okinawa are closely related to the US global strategy	米国の基地は、沖縄の世界戦略と密接に結びついている。/US military bases are closely related to the Okinawa's global strategy
Phrasal Entailment	45	...心をとらえてきた/seized our heart...	...魅了してきた/fascinated...
Polarity	36	...無駄でなかった/not wastefull...	...無駄だった/wastefulness...

Distribution of character-overlap percentage between t_1 and t_2

JA BC dev



JA EXAM dev



There is no guarantee that a good system will result in a comparable performance on a different dataset. This motivates us to build an adaptable/trainable mechanism.

APPROACH

Overview

Given the analysis result, we decided to take a supervised machine learning approach with a careful design on features motivated by linguistics. The classification models we chose are SVM with the linear kernel (in the BC subtasks) and MaxEnt (in the Entrance Exam and RITE4QA subtasks).

Features

Morpheme overlap (through *continuous-to-categorical* conversion); **Basic Element** (Hovy et al., 2006; Fukumoto, 2007) soft-structural overlap; **polarity** (fires when a mismatch of sentiment polarity is captured between t_1 and t_2.); **Morpheme Diff** (takes a diff between t1 and t2, and makes an entailment recognition decision on the different morphemes using character-level heuristic soft-matching); **Quote** – this feature is an N-label indicator, which fires when a quoted content in t_1 occurs in t_2. Intuition: what's written and what's said (or reported in quotation) have different likelihood of being true.

t_1: ...オジサンが「人類は麺（めん）類！」と叫んでいた。/*...the guy was shouting "mankind is a noodle-kind"*
t_2: 人は麺（めん）類だ。/*mankind is a noodle-kind.*

Quantification – fires when there is a mismatch in quantification expression which also indicates an N-label. The quantifier cues: 限って/*only*, 常に/*always* etc

Resources

- **WordNet Synonymy**
- **WordNet Semantic Relatedness –** We **developed and released** a tool that can calculate semantic similarity of two words, using NICT's Japanese WordNet.
- **Wikipedia Hyponymy** – hypernym-hyponym resource automatically extracted from Wikipedia using NICT's Hyponymy extraction tool.
- **Wikipedia Redirect** – We **developed and released** a tool which can extract and utilizes Wikipedia's redirect information to be used for solving alternative forms of the same concept.

Which Features are Useful?

All-but-one feature ablation study

Feature	BC Acc	BC Diff	EXAM Acc	EXAM Diff
All features	62.6%	N/A	68.9%	N/A
- Morpheme Overlap	61.0%	-1.6%	59.1%	-9.8%
- BE Overlap	54.2%	-8.4%	68.9%	0.0%
- Quote	61.4%	-1.2%	68.7%	-0.2%
- Polarity	59.8%	-2.8%	68.7%	-0.2%
- Quantification	62.2%	-0.4%	68.9%	0.0%
- Morpheme Diff	57.2%	-5.4%	68.7%	-0.2%

← lower diff is better. BE has the most negative impact when removed.

# of times features fired

Feature	Y	N
Polarity	30	44
Quote	3	19
Quantification	2	11
Morpheme Diff	15	8

↑ Some features showed high classification power between Y and N

RESULTS

Run	Dev Acc	Test (formal run) Acc	Test (formal run) Top1	Test (formal run) MRR
BC-01	57.6%	53.4%	-	-
BC-02	61.2%	54.2%	-	-
BC-03	*62.6%	54.6%	-	-
EXAM-01	61.1%	60.2%	-	-
EXAM-02	67.5%	65.4%	-	-
EXAM-03	*69.5%	66.7%	-	-
RITE4QA-01	-	84.3%	12.7%	22.2%
RITE4QA-02	-	64.1%	17.4%	25.6%
RITE4QA-03	-	67.5%	21.4%	29.8%

Run 01: BE only baseline.
Run 02: BE + char-overlap baseline.
Run 03: Proposed approach.

* 5-fold cross-validation

APPENDIX: Resources fired on each side of pairs in JA EXAM Dev (showing up to first 10)

WordNet Synonymy
(期/*period*, 時代/*era*), (生まれる/*born*, 誕生/*birth*), (形成/*form*, 結成/*form*), (送る/*send*, 出す/*send*), (発生/, 起こる/*happen*), (商/*commerce*, 貿易/*trade*), (領域/*area*, 地域/*area*), (起きる/*happen*, 起こる/*happen*), (連合/*alliance*, 連盟/*federation*), (議長/*chairman*, 大統領/*President*)

WordNet Semantic Relatedness
(誕生/birth, 生まれる/born), (ポリス/police, 軍/military), (ベール/veil, 保護/protection), (孫/grand-son, 帝/emperor), (行う/execute, 果たす/accomplish), (流通/circulation, 流入/inflow), (生産/production, 構成/composition), (帝/emperor, 始祖/earliest ancestor), (打ち壊す/shatter, 破壊/destroy), (前半/first-half, 後半/last-half)

Wikipedia Hyponymy
(共和国/republic, フィリピン/Philippines), (冤罪/ false accusation, ドレフュス事件/Dreyfus affair), (ロシア連邦大統領/President of Russia, ロシアの大統領/President of Russia), (軍事/military, 軍隊/armed forces), (僧/sangha, 上座/kamiza), (徳川家康/Tokugawa Ieyasu, 徳川氏/Tokugawa clan), (軍事/military, 治安/public safety), (貧困/poverty, 飢饉/famine), (農業/agriculture, 農地/arable land), (福祉/welfare, 介護/elderly care)

Wikipedia Redirect
(アナトリア半島/Anatolia, 小アジア/Asianic), (ギリシャ/Greece, ギリシア/Greece), (1/1, 第一/first), (平和条約/peace treaty, 講和/pacification), (ヴァスコ・ダ・ガマ/Vasco da Gama, ヴァスコ＝ダ＝ガマ/Vasco da Gama), (イスラム/Islam, イスラーム/Islam), (イスラム教/Islam, イスラーム教/Islam), (マリ帝国/Mali Empire, マリ王国/Mali Empire), (アメリカ人/American, アメリカ合衆国/U.S.A.), (アパルトヘイト/apartheid, 人種隔離政策/racial segregation policy),

CONCLUSION

We presented the LTI's system participated in NTCIR-9 RITE. Through analysis, we assumed that multiple linguistic phenomena must be captured, and there is a need of adaptability in a Textual Entailment recognition system. We experimentally showed that they are reasonable assumptions to make. Tools we built, i.e. WS4J[1] and Wikipedia Redirect[2], are released as open source software.

[1] http://code.google.com/p/ws4j/
[2] http://code.google.com/p/wikipedia-redirect/

Future work: elaborate more on capturing linguistic modalities. Especially, recognizing *epistemic modality*, or *committed and non-committed belief* could be a sophisticated extension of the *Quote* feature we used.