RTE-5@TAC2009

The Fifth Recognizing Textual Entailment Challenge

Luisa Bentivogli (coordinator, CELCT & FBK-irst)
Danilo Giampiccolo (coordinator, CELCT)
Hoa Trang Dang (NIST)
Ido Dagan (Bar Ilan University)
Bernardo Magnini (FBK-irst)

Outline

- The RTE Challenge
- RTE-5 Main Task
  - RTE*1to5*: Main Task Evolution
  - Knowledge Resources for RTE
- RTE-5 Pilot Search Task
- Conclusion and Future Perspectives

Outline

- The RTE Challenge
- RTE-5 Main Task
  - RTE*1to5*: Main Task Evolution
  - Knowledge Resources for RTE
- RTE-5 Pilot Search Task
- Conclusion and Future Perspectives

Textual Entailment

Textual entailment is a directional relation between two text fragments:

- the entailing text, called ***T***(ext)
- the entailed text, called ***H***(ypothesis)

T entails H if, typically, a human reading T would infer that H is most likely true

Task(s) Definition

Given T and H systems must decide whether:

- **2-way task:**
  - T entails H (*ENTAILMENT*)
  - T does not entail H (*NO ENTAILMENT*)
- **3-way task:**
  - T entails H (*ENTAILMENT*)
  - T contradicts H (*CONTRADICTION*)
  - The truth of H cannot be determined on the basis of T (*UNKNOWN*)

Examples

- **YES**

  T: A shootout at the Guadalajara airport in May, 1993, killed Cardinal Juan Jesus Posadas Ocampo.

  H: Juan Jesus Posadas Ocampo died in 1993.

- **CONTRADICTION**

  T: Seven miners have been killed after a coal mine flooded in north China.

  H: A coal mine accident killed more than 73 people in China.

- **UNKNOWN**

  T: 632 Air Canada flight attendants will lose their jobs in November.

  H: European Airlines are cutting jobs.

The RTE-5 Challenge

- Proposed for the second time as a track at the Text Analysis Conference (TAC2009) organized by NIST
- Main Task structure remained unchanged
  - traditional two-way task
  - three-way task introduced in RTE-4
- Ablation tests on knowledge resources used by systems participating in the Main task
- Pilot Search task situated in the Summarization application setting

RTE-5 Participants

- Number of participants: **21**
  - RTE-1: 18, RTE-2: 23, RTE-3: 26, RTE-4: 26
- Provenance
  - NORTH AMERICA: 5, SOUTH AMERICA: 1, EU: 8, ASIA: 5, AUSTRALIA: 2
- Participants per task
  - Main Task: 20 (54 runs)
  - Pilot Search Task: 8 (20 runs)

Outline

- The RTE Challenge
- RTE-5 Main Task
  - RTE*1to5:* Main Task Evolution
  - Knowledge Resources for RTE
- RTE-5 Pilot Search Task
- Conclusion and Future Perspectives

The Data Set

- Development Set and Test Set
- T-H pairs: 1,200 (600 Dev Set + 600 Test Set)
- Application settings
  - IE (200+200), IR (200+200), QA (200+200)
  - NO SUM
- Distribution wrt the entailment judgment:
  - 50% *YES*, 35% *UNKNOWN*, 15% *CONTRADICTION*
- Longer *T's (100 words vs. 40 words in* RTE-4)
- T's not edited from their source documents

Evaluation Measures

Automatic evaluation:

- **Accuracy** *(main evaluation measure)*: percentage of correct judgments against the Gold Standard
- **Average Precision** *(for systems which returned a confidence-ranked list of the test set pairs):* average of the system's precision values at all points in the ranked list in which recall increases, that is at all points in the ranked list for which the gold standard annotation is YES.

Participants

- Teams: 20
  - 3-way task only: 7
  - 2-way task only: 10
  - Both tasks : 3
- Runs
  - 3-way task: 24
  - 2-way task: 54
    - 30 explicitly submitted to the 2-way task
    - 24 derived from the 3-way runs

Results: Accuracy Statistics

	3-way Task		2-way Task	
	All runs	Best runs	All runs	Best runs
Highest	68.33	68.33	73.5	73.5
Lowest	43.83	46.83	50.00	50.00
Median	52.00	55.83	61.08	61.5
Average	52.91	56.1	60.36	61.52

Results: RTE-5 vs. RTE-4

	3-way Task		2-way Task	
	All runs	Best runs	All runs	Best runs
Highest	68.33	68.33	73.50	73.50
RTE-4	68.50	68.50	74.60	74.60
Lowest	43.83	46.83	50.00	50.00
RTE-4	30.70	30.90	49.70	51.60
Median	52.00	55.83	61.08	61.50
RTE-4	54.30	55.00	57.05	58.30
Average	52.91	56.10	60.36	61.52
RTE-4	50.65	52.59	58.03	59.41

Results: Accuracy per Task

Task	3-way Task	2-way Task
	Average Accuracy	Average Accuracy
IE	47.25	53.31
QA	51.15	57.45
IR	60.33	70.32

Best Results

3-way Task		2-way Task	
Run	**Accuracy**	**Run**	**Accuracy**
UAIC20091	0,6833	UAIC20091-3*way*	0,735
DFKI2	0,6367	DFKI3-3*way*	0,685
DLSIUAES1	0,600	QUANTA1	0,670
AUEBNLP2	0,575	PeMoZa2	0,6617
rhodes1	0,570	UI_ccg1	0,6433
Boeing3	0,5467	BIU2	0,6383
cswhu1	0,5217	cswhu1-3*way*	0,6333
Sagan1	0,5217	DLSIUAES2	0,6317

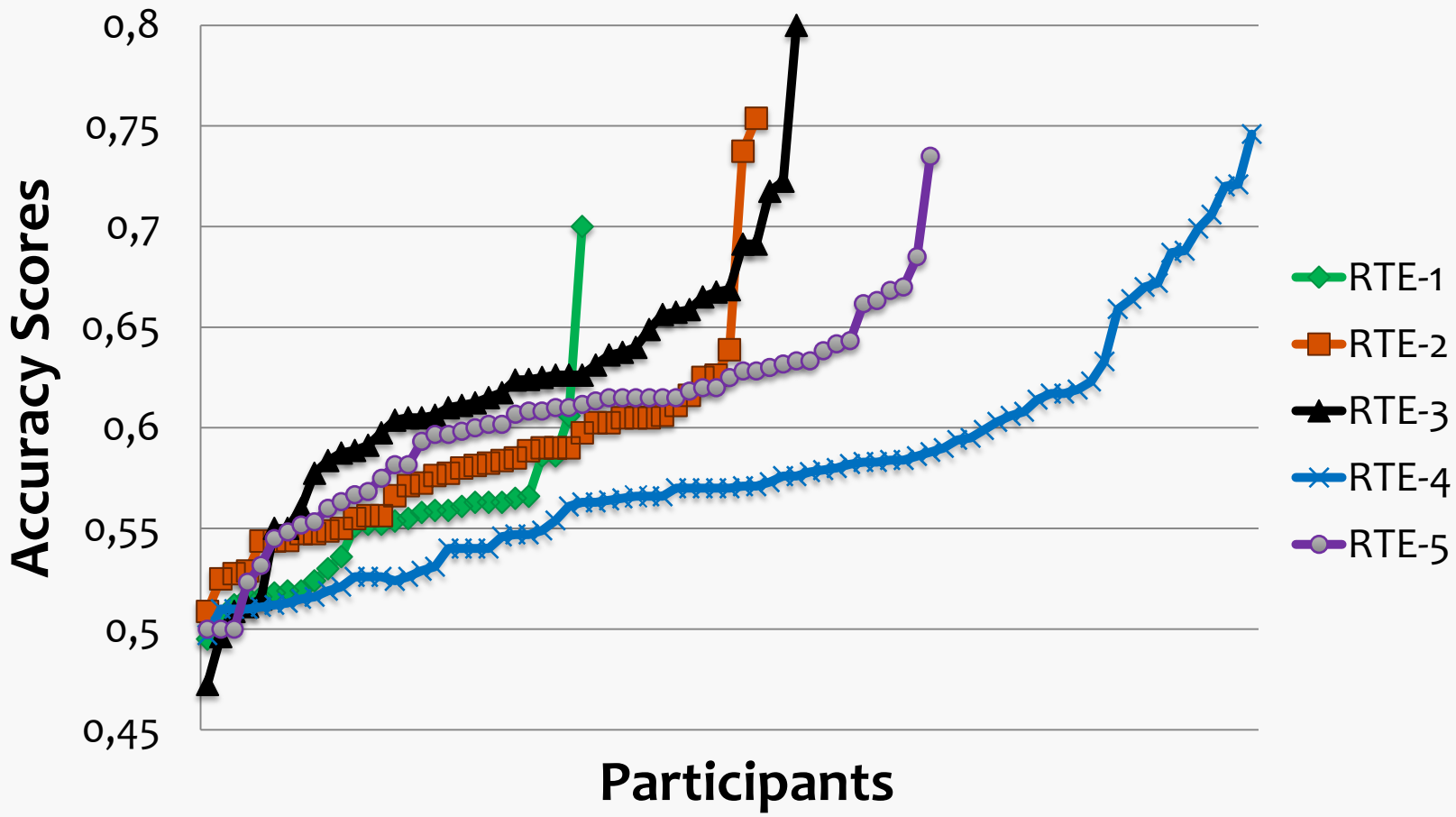
Outline

- The RTE Challenge
- RTE-5 Main Task
  - RTE*1to05:* Main Task Evolution
  - Knowledge Resources for RTE
- RTE-5 Pilot Search Task
- Conclusion and Future Perspectives

RTE-1 *to* 5 Datasets

Challenge	Data Set	# of Pairs	H length (# of words)	T length (# of words)
RTE-1	DEV	567	10,08	24,78
	TEST	800	10,8	26,04
RTE-2	DEV	800	9,65	27,15
	TEST	800	8,39	28,37
RTE-3	DEV	800	8,46	34,98
	TEST	800	7,87	30,06
RTE-4	TEST	1000	7,7	40,15
RTE-5	DEV	600	7,79	99,49
	TEST	600	7,92	99,41

RTE-1 *to* 5 Results: 2-way Task



RTE-*1 to 5* Baselines

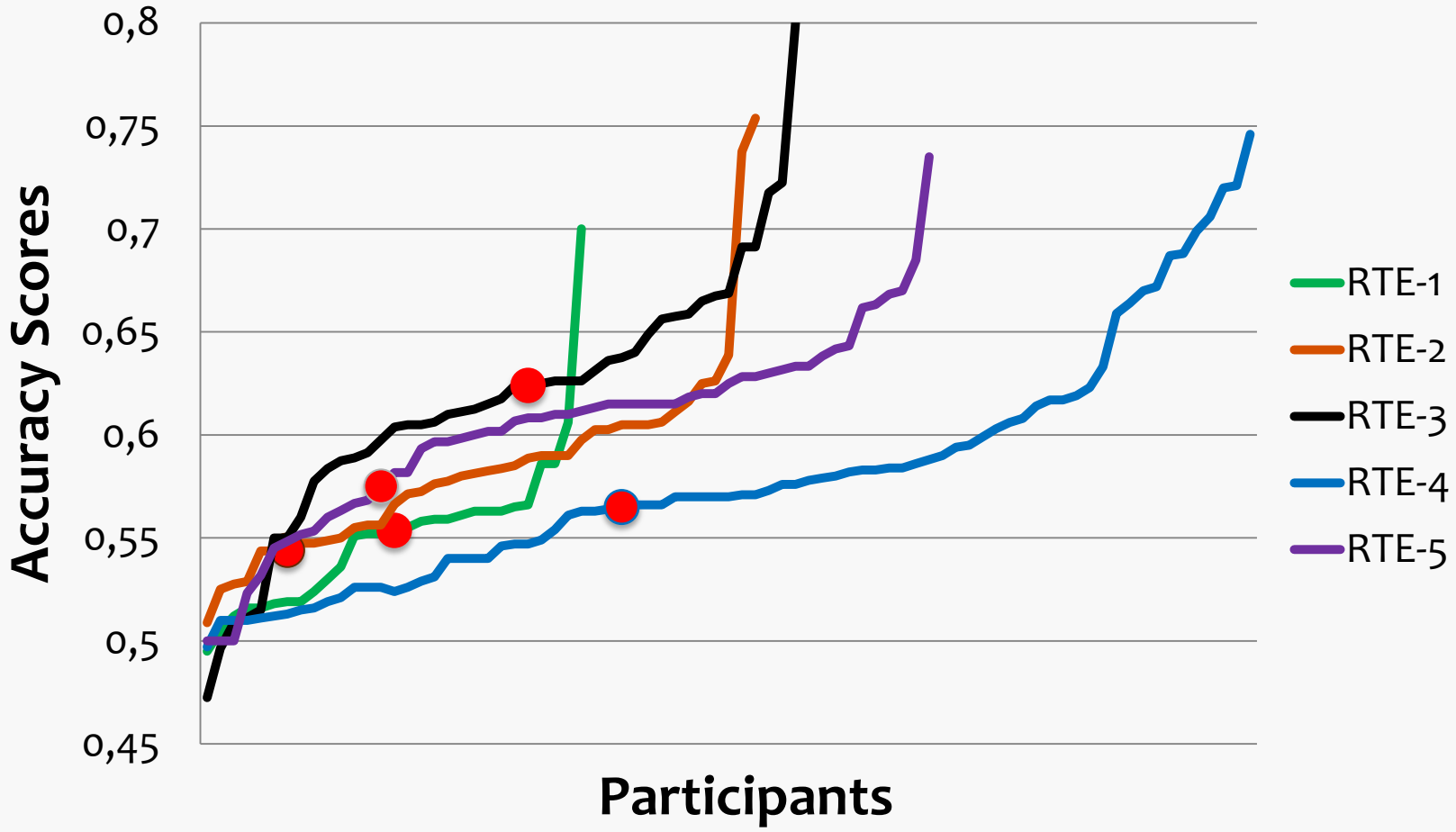
Mehdad and Magnini (2009)[1]

- Word overlap baseline
  - Pre-processing: *TreeTagger*[2]
  - T/H overlap: *Text Similarity* package[3]
  - Classification: TinySVM package[4]
  - 8 different settings (lemma|tokens, overlap normalization, stopwords)

1. http://hlt.fbk.eu/sites/hlt.fbk.eu/files/baseline.pdf
2. *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/*
3. *http://www.d.umn.edu/~tpederse/text-similarity.html*
4. *http://chasen.org/~taku/software/TinySVM/*

RTE-1 *to* 5: 2-way Task Results with Baselines

Baseline setting # 8: H/T tokens, no stopwords, no normalization

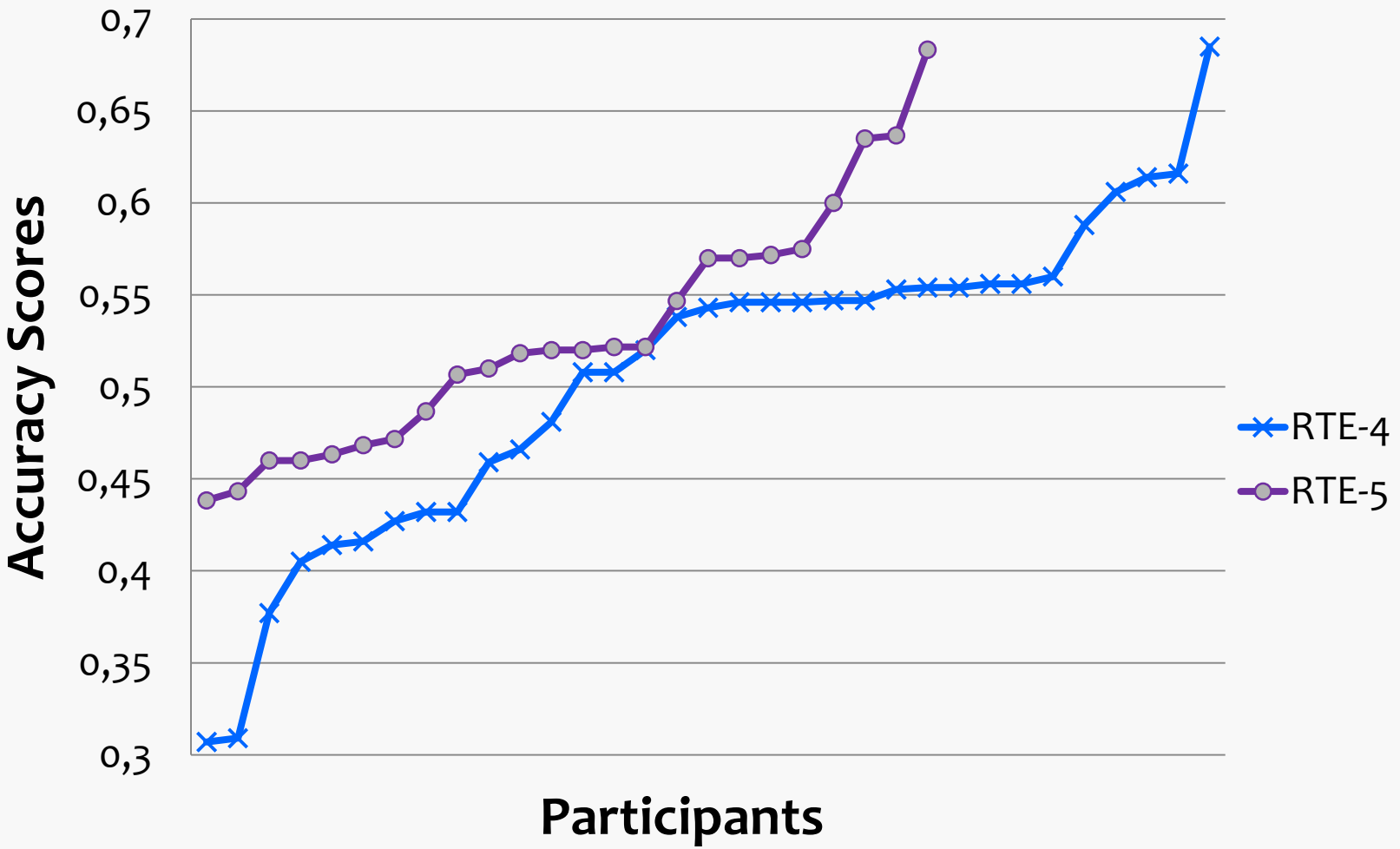


RTE-*1 to 5* Task Difficulty

Baseline setting # 8: H/T tokenization, stopwords excluded, overlap not normalized

BASELINE		RTE TEST SETS			
SETTING # 8		H/T OVERLAP (%)			OVERLAP DIFFERENCE between YES and NO/UNKN
TASK	ACCURACY (%)	CONTRADIC	YES	NO ENTAIL/ UNKNOWN	
RTE-1	55.37		68.64	64.12	4.52
RTE-2	54.4		70.63	63.32	7.31
RTE-3	62.4		69.62	55.54	14.08
RTE-4	56.6	67.97	68.95	57.36	11.59
RTE-5	57.5	78.93	77.14	62.28	14.86

RTE-1 *to* 5 Results: 3-way task



Outline

- The RTE Challenge
- RTE-5 Main Task
  - RTE-*1to5* Main Task Evolution
  - Knowledge Resources for RTE
- RTE-5 Pilot Search Task
- Conclusion and Future Perspectives

Knowledge Resources for RTE

A new initiative aimed at studying the relevance of knowledge resources in recognizing TE

- ***Ablation Tests*** for all knowledge resources used in Main Task runs:
  - remove one module at a time from a system, and re-run the system on the test set with the other modules, except the one tested

  ! Remove only knowledge resources

  ! Remove one resource at a time

Ablation Tests

- 82 ablation tests submitted (by 19 teams)
  - 29 tests did not specifically address knowledge resources (e.g. pre-processing modules, entailment algorithms, estimated thresholds, statistical features)
  - In 16 tests a combination of different resources/components was removed from the system instead of one single resource
- 37 ablation tests conformant to the requirements

Ablation Tests Results: 2-way Task

Ablated Resource	# of Ablation Tests	Impact on Systems		
		Positive	Null	Negative
WordNet	19	9 (+1.48%)	3 (--)	7 (-0.71%)
VerbOcean	6	2 (+0.25%)	3 (--)	1 (-0.16%)
Wikipedia	4	3 (+1.17%)	0	1 (-1%)
FrameNet	3	1 (+1.16%)	1 (--)	1 (-0.17%)
DIRT	3	2 (+0.75%)	0	1 (-1.17%)
PropBank	1	1 (+2%)	0	0
Acronym-guide	1	0	1 (--)	0
Total	37	18	8	11

Ablation Tests Results: 3-way Task

Ablated Resource	# of Ablation Tests	Impact on Systems		
		Positive	Null	Negative
WordNet	9	4 (+1.71%)	1 (--)	4 (-1%)
VerbOcean	4	3 (+0.28%)	1 (--)	0
Wikipedia	2	2 (+2.42%)	0	0
FrameNet	1	0	0	1 (-0.17%)
DIRT	2	1 (+0.33%)	1 (--)	0
PropBank	1	1 (+3.17%)	0	0
Acronym-guide	1	0	1 (--)	0
Total	20	11	4	5

The Top Impact Resource

- Positive impact: WordNet
  - Boeing 3-3*way* *(synonyms, hypernyms, similar,pertains, derivational)*
    - 3-way evaluation: 5.67%
    - 2-way evaluation: 4%
  - UI_ccg1 *(word similarity == identity)*
    - 2-way evaluation: 4%
- Negative impact: WordNet
  - AUEBNLP1-3*way* *(synonyms)*
    - 3-way evaluation: 2.67%
    - 2-way evaluation: 2%

Lesson Learned

- Definition of knowledge resource not clear cut (e.g. Named Entities, stopword lists, negation rules, ...)
- Determining the actual impact of knowledge resources is not straightforward
  - Different uses -> different impacts
- Need for a deeper comprehension of the usage of the resources
- Effort towards normalization: try to individuate the best way to use the knowledge contained in the resources

Outline

- The RTE Challenge
- RTE-5 Main Task
  - RTE*1to5*: Main Task Evolution
  - Knowledge Resources for RTE
- RTE-5 Pilot Search Task
- Conclusion and Future Perspectives

The *RTE-5* Search Pilot Task

Motivation:

- Move towards more realistic scenarios: test RTE systems against real data
- Analyze the potential impact of entailment in a real NLP application scenario like SUM

The *RTE-5* Search Pilot Task

- Systems must find all the sentences that entail a given H in a given set of documents about a topic
- Summarization application setting:
  - **H's** are based on Summary Content Units that have been created from human-authored summaries for a corpus of documents about a common topic
  - **T's**, i.e. the entailing sentences, are to be retrieved in the corpus for which the summaries were made

The *RTE-5* Search Pilot Task

H's SET

H1: The AS-28 mini-submarine was trapped underwater

H2:Seven submariners were onboard the AS-28

H4: Russia requested international help to rescue the AS-28

H5:The AS-28 crew was rescued in satisfactory conditions

Document 1

S1: Effort seen to raise stricken Russian sub

S2: The effort to attach the cables marked the start of an operation to raise the vessel trapped below the surface with seven crewmen on board.

task," Fyodorov said.

…

S9: Fyodorov said earlier that the seven crewmen were in satisfactory condition...

…

Document 2

S0: Japan sends help to trapped Russian submarine

S1: Japan on Friday dispatched four military ships to help Russia rescue seven crew members aboard a small submarine trapped on the seabed in the Far East.

spokesman said.

S5: "We will do our utmost efforts to rescue them.

S6: We are hopeful," he said.

S7: The assistance comes despite rocky relations between Japan and Russia, which have yet to formally end World War II amid Japan's claims to four islands off Hokkaido that Soviet troops seized in August 1945.

…

Document 3

S0: Russian sub snagged on undersea surveillance antenna: official

S1: Rescue of a submarine stuck on the seabed off Russia's east coast is complicated because it is snagged in an underwater surveillance antenna system as well as snared in a fishing net, a senior Russian naval officer said Friday.

S2: A remote-controlled device was lowered to the stricken vessel "to cut the flexible tubes and cables of the coastal surveillance antenna in which the AS-28

S6: There are seven crew members aboard the vessel, stranded on the ocean floor in a bay off the coast of the Kamchatka peninsula in Russia's Far East region.

Main vs. Pilot Task

Main Task

- Classification task
- The distribution of entailment is determined a priori
- T and H are artificially created and do not contain references to information outside the pair itself

Search Task

- Retrieval task
- Reflects the natural distribution of entailment in a corpus
- Both T and H are to be interpreted within the context of the topic, as they rely on explicit and implicit references to entities, events, dates, places, etc. pertaining to the corpus

Data Set Description

- Data taken from the TAC Update Summarization task:
  - Development Set: SUM 2008
  - Test Set: SUM 2009
- For each Topic:
  - a corpus of 10 newswire documents
  - between 6 and 10 Hypotheses
- All documents manually split into sentences, which represent the T's to be judged for entailment

Data Set Composition

DEVELOPMENT SET		TEST SET	
Topics	10	Topics	9
Hypotheses	80	Hypotheses	81
Sentences	2,538	Sentences	1,949
Annotations	20,104	Annotations	17,280
"entailing" judgm.	810	"entailing" judgm.	800

- 3 annotations for the whole data set
- IAA *(Kappa)*: 97.10% (Dev), 97.02% (Test)

Evaluation

8 participants (20 runs)

Evaluation measures:

- Precision, Recall, F-measure
  - Micro-averaged: official metrics
  - Macro-averaged
    - by Topic and by Hypothesis
    - If no sentence is returned for a given Topic/Hypothesis, Precision for that Topic/Hypothesis is set to *0*

Results: F-measure statistics

Micro-averaged results:

F-measure	All runs	Best runs
Highest	45.59	45.59
Lowest	9.55	17.51
Median	30.14	30.2
Average	29.17	30.51

Search Task Baseline

Information Retrieval baseline:

- Each Topic is a corpus
- Sentences are "the documents" to be retrieved
- Hypotheses are the queries
- LUCENE text search engine:
  - *StandardAnalyzer* (tokenization, lower-case and stop-word filtering, basic clean-up of words)
  - Boolean "OR" query
  - Default Lucene ranking
  - Select the top-ranked (5, 10, 15, 20) sentences

Best Results

Micro-averaged results:

Team	Precision	Recall	F-measure
Baseline_10	*0,4691*	*0,475*	*0,472*
BIU3	0,4098	0,5138	0,4559
unimelb1	0,4294	0,38	0,4032
FBKirst2	0,2254	0,6475	0,3344
UAIC20092	0,5112	0,2288	0,3161
clr092	0,2034	0,4925	0,2879
Boeing2	0,3339	0,2512	0,2867
Sagan1	0,1016	0,855	0,1816
ssl1	0,1149	0,3675	0,1751

Outline

- The RTE Challenge
- RTE-5 Main Task
  - RTE*1to5*: Main Task Evolution
  - Knowledge Resources for RTE
- RTE-5 Pilot Search Task
- Conclusion and Future Perspectives

Conclusions

- Main Task
  - average performances increased
- Evaluation of Knowledge Resources
  - very posite response
  - first step towards sharing and reuse of resourses
- Pilot Search Task
  - interaction between the RTE and SUM tasks
  - textual entailment recognition performed on a real corpus
  - natural distribution of entailment

Future Directions

See you all at the RTE Planning Session

Thank you!