Using MT-Based Metrics for RTE

Alexander Volokh
and
Günter Neumann
DFKI, Germany

German Research
Center for Artificial
Intelligence GmbH

Result Analysis

- RTE-6 results:
  - 38 - 48 F1-score
  - Good or bad?
- What is the highest realistic result within the scope of RTE-7?
  - What are the different phenomena in the data?
  - How frequent?
  - How difficult?

Complexity Analysis

- Divided the data into three complexity classes:
  - A: syntax
  - B: lexical semantics (synonymy)
  - C: inference / world knowledge

  → A – 30% of the data, B – 35%, C – 35%
- Focus on A and B classes, C – too difficult for the scope of the task.

Examples (A class)

- H_1: People were forced to leave their pets behind when they evacuated New Orleans.
- T_1: Thousands of people were forced to leave their pets behind when they evacuated New Orleans.

- the relevant information is expressed with the same words in both T and H
- analysis of the syntactic structure should suffice

Examples (B class)

- H_1: People were forced to leave their pets behind when they evacuated New Orleans.
- T_2: Animal rescue officials have been collecting scores of pets and other animals from the shattered city, while many survivors have told of their distress at having to leave beloved cats and dogs behind in the watery city when they fled.
- T_3: Such emotional scenes were repeated perhaps thousands of times along the Gulf Coast last week as pet owners were forced to abandon their animals in the midst of evacuation.

- the words used in T_2 and T_3 differ from those used in H_1
- one has to know about the synonymy/semantic relatedness of the words in addition to the syntactic structure

Examples (C class)

- H_1: People were forced to leave their pets behind when they evacuated New Orleans.
- T_4: For Elizabeth Finch, the owner of two dogs named Zorra and Hans Blix, the sight of citizens forced to choose between their pets and their safety was, like the disaster itself, indicative of broader social rifts.
- T_5: The animals are being cared for at a farm north of Louisiana until they can be reunited with their families, many of whom were told they would not be able to bring their pets on evacuation busses and helicopters.

- Logic inference and/or world knowledge

Approach for A and B

- A and B – same or similar words are used
- Idea: assume T and H are translations of the same source sentence

  → The assumption is wrong in general, T contains more information than H (in YES case)
- But: the similarity between T and H in YES cases is still higher than in NO cases
- → Entailment can be predicted

Meteor

- We compute the similarity using different features
- Most important ones use Meteor[1] (Metric for Evaluation of Translation with Explicit Ordering)
- Meteor matches words using *exact*, *stem* and *synonym* modules
- If T entails H Meteor score is higher than if T does not entail H (especially for A, but also for B classes)

1- http://www.cs.cmu.edu/~alavie/METEOR/

Weaknesses

- Problems:

  - Does not work if T and H have completely different lengths

  - Synonym module does not always match

  - Finally, T is simly not equal H

- Final result far below the targeted boundary of 0.65

Conclusion

- 43.41 micro-average F1-score
- 46.34 macro-average F1-score

  → Above median, big improvement over the last year
- Very robust solution to an exremely large amount of data

  - >50% can be solved this way if account for weaknesses
- Problem-specific alternatives can still be included for the rest of the data