Detecting Irrelevant Sentences in Mathematical Word Problems

Introduction and Motivation

- Mathematical word problems is one of the most popular tools to introduce mathematics
- Consists of a narrative
- Usually has 3-8 sentences, followed by a question at the end
- Irrelevant sentences and numbers in the narrative

Introduction and Motivation

- “The Library of Congress has the largest comic book collection in the United States. It contains about 100,000 pieces. It is growing by about 200 issues each month. About how many more issues will the library have in 3 months.”

- First and second sentences are irrelevant

Methods

- Independent Prediction
  - Support Vector Machine
    - SVMLight
  - Bayesian Logistic Regression
    - BBR
- Joint Prediction
  - Boltzmann machine

Independent Prediction

- Create a model for SVM and BBR using training data
- Learn the threshold
- Use SVM and BBR to classify

Independent Prediction

- Improvement to the first step can be made
- Use the sim(Q,Si) as a feature
- Similary between two sentences:

$$Weight(q_t, d_t, Q, D) = term\_freq(d_t) * \log\left(\frac{totalNumDocs}{numDocsContain(termt) + 1}\right) * Weight(q_t)$$

Joint Prediction

$$P(S_1, S_2, \ldots, S_N) = \frac{1}{Z}$$
$$* \exp\Big(\sum_i \alpha_1 SVM(S_i) * \delta(S_i == 1)$$
$$+ \sum_i \alpha_0 \delta(S_i == 0) + \sum_i \sum_j \beta_{11} Sim(S_i, S_j) * SVM(S_i) * SVM(S_j) * \delta(S_i == 1)$$
$$* \delta(S_j == 1) + \sum_i \sum_j \beta_{10} Sim(S_i, S_j) * SVM(S_i) * SVM(S_j) * \delta(S_i == 1)$$
$$* \delta(S_j == 0) + \sum_i \sum_j \beta_{01} Sim(S_i, S_j) * SVM(S_i) * SVM(S_j) * \delta(S_i == 0)$$
$$* \delta(S_j == 1) + \sum_i \sum_j \beta_{00} Sim(S_i, S_j) * SVM(S_i) * SVM(S_j) * \delta(S_i == 0)$$
$$* \delta(S_j == 0)$$

Evaluation Methodology

- F1 = 2*p*r / (p+r)
  - p is precision
  - r is recall

  for irrelevant sentences

Data Set

- The data set consists of 92 questions :
  - 9 questions with 1 irrelevant sentence and 0 irrelevant number
  - 6 questions with 2 irrelevant sentences and 0 irrelevant number
  - 3 questions with 3 irrelevant sentences and 0 irrelevant number
- …
- **Each of the question has tags to indicate the irrelevant sentences and numbers.**

Data Set

- <irrel_S>Luis and his family drove from New York City to the Grand Canyon. <irrel_S>
- <irrel_S>The rent a car for <irrel_N>$50<irrel_N> per day<irrel_S>.
- The drove for a total of 50 hours at an average speed of 50 miles per hour.
- About how far did they drive?
- (Grade4, p. 236, Learn)

Result: Step 1

BBR

CONFIGURATION	BEST_F1
with_all	0.608108108108101

SVM

CONFIGURATION	BEST_F1
with_all	0.631578947368421

Result: Step 2

BBR

CONFIGURATION	BEST_F1
with_all	0.721804511278195

SVM

CONFIGURATION	BEST_F1
with_all	0.75

Reference

- [1] K. Jeongwoo, S. Luo, and N. Nyberg. A probabilistic Graphical Model for Joint Answer Ranking.2007.
- [2] G. Alexander, L. David, and M. David. Large-Scale Bayesian Logistic Regression for Text Categorization. 2006.
- [3] J. Thorsten. Learning to Classify Text Using Support Vector Machines. 2002.

Question?