Design and Engineering of Intelligent Information Systems (DEIIS)

Homework #1

Weston Feely
andrew ID: wfeely
email: wfeely@cs.cmu.edu

September 11, 2013

Type System Description

My type system includes the following types.

0.1 BaseAnnotation

The base annotation type required by the assignment description. This includes two features.

Features:

- string source: string to help keep track of where an annotation was originally made.
- double confidence: how confident the annotation was.

0.2 Question

The question type for representing questions from the input data. This includes six features.

Features:

- int begin: the beginning character offset for the question string.
- int end: the ending character offset for the question string.

- FSArray(Token) tokens: a Token array of tokens, for representing words from the question string.
- FSArray(Ngram) unigrams: a Ngram array of bigrams, for representing 1-token strings from the question string. (similar to tokens).
- FSArray(Ngram) bigrams: a Ngram array of bigrams, for representing 2-token strings from the question string.
- FSArray(Ngram) trigrams: a Ngram array of trigrams, for representing 3-token strings from the question string.

0.3 Answer

The answer type for representing answers from the input data. This includes eight features.

Features:

- int begin: the beginning character offset for the answer string.
- int end: the ending character offset for the answer string.
- FSArray(Token) tokens: a Token array of tokens, for representing words from the answer string.
- FSArray(Ngram) unigrams: a Ngram array of bigrams, for representing 1-token strings from the answer string. (similar to tokens).
- FSArray(Ngram) bigrams: a Ngram array of bigrams, for representing 2-token strings from the answer string.
- FSArray(Ngram) trigrams: a Ngram array of trigrams, for representing 3-token strings from the answer string.
- boolean gold: boolean for representing the 1 or 0 from the answer input text, which represents whether this is a good answer or a bad answer.
- boolean guess: boolean for representing the guess (0 or 1) for whether this answer is good (1) or bad (0), based on the other features (excluding the gold variable).

0.4 Token

Token type for representing the words in a sentence. This includes three features.

Features:

- int begin: the beginning character offset for the token.
- int end: the ending character offset for the token.
- string token: the string for the word token, itself.

0.5 Ngram

Ngram type for representing the ngrams (space-separated strings of words) in a sentence. This includes four features.

Features:

- int order: the order of the ngram represented (1 for unigram, 2 for bigram, 3 for trigram).
- int begin: the beginning character offset for the ngram.
- int end: the ending character offset for the ngram.
- string ngram: the string for the ngram, itself.

0.6 Eval

Evaluation type for representing evaluation objects. These objects wrap together the answers, and their corresponding question, along with the features necessary to perform scoring and ranking of answers, as well as the final performance evaluation for a given input data set. This includes five features.

Features:

- Question question: question object for the question in this data set.
- FSArray(Answer) answers: answer object array for the answers in this data set.
- doubleArray performance: scores for each answer in this data set.

- int n: the number of correct answers in this data set.
- double p: the precision for this data set (# answers guessed correct divided by the number of total correct answers in this data set, n).