Textual Entailment – Fitchburg State College

Orlando Montalvo-Huhn* Stephen Taylor †

October 17, 2008

Abstract

Our submission guesses at entailment based on word similarity between the hypotheses and the text. We attempt three kinds of comparisions: original words (with normalized dates and numbers) synonyms, and antonyms. Each of the three comparisions contributes a different weight to the entailment decision.

Our results are insignificantly better than chance for the two-way comparison. However, for the three-way comparison they are much better.

1 Introduction

Our group spent a long time on parsing, with the intention of deriving Basic Elements from the sentence. We planned to attempt logical inference, and perhaps compute a score for approximate inferences. However, as the submission deadline approached, we backpedaled, and modified one of the modules from our question-answering submission of last year.

Our submission guesses at entailment based on word similarity between the hypotheses and the text. We attempt three kinds of comparisions: original words, synonyms, and antonyms. Each of the three comparisions contributes a different weight to the entailment decision.

2 Original word comparison

Stop-words are removed from the wordlist, and repetitions of a word are ignored.

We have a series of regular expressions which match different kinds of dates, times, and numbers. A normalized date is represented as Month day, year CE.

Thus the dates Sep 1 '08 and September 1, 2008 would both be represented as September 1, 2008 CE.

Each word in the sentence is weighted according to its inverse frequency, as recorded in our database (derived from the 2007 QA

*Andy's address at IBM: Orlando_Montalvo-Huhn@us.ibm.com

†Steve's address at Fitchburg State College: staylor@fsc.edu

training data.)

Words for which the frequency is unknown are a special case, in that we know from the fact that they occur in the query that the frequency count should not be zero, even though we didn't observe the word in our run through the frequency corpus. For these frequencies we use Witten-Bell discounting, following section 6.3 of Jurafsky and Martin[1]. We assume that the probability mass of never-observed words is:

$$\sum_{c_i=0} p_i^* = \frac{T}{N+T}$$

where T is totalTypes ever seen, and N is totalTokens.

The word-vector W_i with weights w_i for the hypotheses is compared to the word-vector V_j, with weights v_j for the text, using a cosine similarity measure:

$$m_1 = \frac{\sum_{V_i=W_j} w_i * v_j}{\sqrt{(\sum_k w_k^2)\,(\sum_k v_k^2)}} \qquad (1)$$

3 Synonym matching

For synonym matching, two words match if they are both members of the same Wordnet synset. They may also match according to the following strategy: If a word has less than five synonyms (total members of all synsets it participates in) we will also match against direct hypernyms. If there are still less than five words in all synsets, we consider also direct hyponyms. Thus for each word, there are several synsets which may be synsets of the original word, or synsets of its hypernym or hyponym, against which the word can be matched.

The weight of each synset is computed by using wordnet frequency counts. These are available for each sense. Each W_i has has its frequency w_i, as above, but in addition, W_i participates in n_i synsets, each synset has m_{n_i} senses, and a frequency c_{n_i}. The adjusted weight for the synset is

$$\frac{c_{n_i}}{\sum_{1\le r\le n_i} m_r} w_i$$

Thus the weights for a match might be different on the left and right sides of the match.

However, as with the original words, there may be many more synsets in the text than in the the hypotheses.

4 Antonym matching

Antonym matching uses the same algorithm as synonym matching, but the synsets are the Wordnet antonyms.

5 Two-way answers

For the two-way algorithm, the possible answers are drawn from the set: $\{\text{entailed}, \text{not entailed}\}$

We compare the original words and the synonyms from the text and the hypothesis, using the algorithms described above, and if the value is above a threshold, we pronounce the hypothesis entailed, otherwise not. We learned the threshold by maximizing our score on a subset of last year's training data.

Synonyms entailed	antonyms entailed	three-way answer
no	no	unknown
no	yes	not entailed
yes	no	entailed
yes	yes	unknown

Table 1: Three-way decision table

6 Three-way answers

For the three-way algorithm, the possible answers are drawn from the set: {entailed, not entailed, unknown} Our algorithm for three-way decisions uses our two-way algorithm twice: first it computes a score for matching against the original words and the synonyms. Then it computes a score for matching antonyms of words from the hypothesis against the text. The results given are based on the four possibilities for entailment as shown in table 1

7 Results

For the two-way comparison, we scored 52.6% correct. For the three-way comparison, our score was 46.6% correct.

Assuming 1000 items with 50/50 probability, the standard deviation of the expected score is

$$\sqrt{(.5)(.5)1000} \approx 16.3 \text{ items}$$

or 1.63 percent.

Our result of 526 items correct is 1.6 standard deviations above the expectation for the mean. To be significant at the 5% level would require 1.645 standard deviations.

A similar calculation for the three-way results gives much better significance. Since there are three possible answers, the chance of hitting one by chance should be 1/3. The expected score is 33.3%, and the standard deviation of the expected score is

$$\sqrt{(.333)(.667)1000} \approx 14.9 \text{ items}$$

or 1.49 percent. That makes our score of 46.6% or $\mu + 8.93\sigma$ wildly significant.

Perhaps the result is an artifact. It may be that our algorithm and the data both happen to favor an "unknown" answer. However, our algorithm for the three-way result is different than our algorithm for two-way results. It looks possible that a weak algorithm like our two-way algorithm can be combined to give much better results in the three-way decision.

References

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice-Hall, 2000.

[2] George A. Miller. http://wordnet.princeton.edu/.