The Institute for Informatics,
Logics and Security Studies

Comparing Corpora for Classifier Disambiguation

Patrick Zongo, Samira Shaikh

12/01/2009

ILS Tech Report: 009

Available for download from:
`http://www.ils.albany.edu`

Abstract

In information retrieval the problem of being able to efficiently classify data is ubiquitous. The following analysis will extend earlier research in comparison of n-gram frequency in order to annotate communicative actions.

1 Introduction

Comparing corpora has been used in previous research in order to determine words which are specific to a field[JSE+05]. Implementation of this method compared a sample corpus with a known general corpus. This allows for a system to generate terms which can be expected to be found in the sample corpus.

In our analysis, we wanted to verify that corpora comparison could be modified for the task of automatic communicative act annotation. If it can, we can extend the notion of corpora comparison to solve various problems in natural language processing.

2 Data Collection

The data that were collected in this project are seven online chat discussions. These discussions contain approximately 400 utterances each of task-oriented conversation. For our purposes we define a task as a sequence of utterances which results in the selection of one option out of a group of choices. One example of task-oriented chat required the speakers in the chat room to choose the recipient of a job from a set of applications.

After the data were collected, annotators determined which communicative action were assigned to each utterance. The communicative actions we dealt with were *continuation-of*, *addressed-to*, and *response-to*. The meanings for these actions are defined below:

Continuation-of This annotation indicates that the given utterance was a continuation of a previous utterance.

Addressed-to When an utterance is labeled as an *addressed-to* it indicates the utterance was directed at some other speaker.

Response-to When an utterance is labeled as a *response-to* it indicates that the speaker was responding to another speaker's previous utterance.

To develop a ground truth to train our system on. We iterated through each utterance and labeled it as the communicative action that the majority agreed on.

3 Corpora Comparison

In two completely identical corpora, the frequency of an n-gram in one corpus is identical to the frequency of the n-gram in the other. However over two different corpora, an n-gram that has a high frequency in one corpus could have a lower frequency in the other corpus. The n-grams that have great disparity over two corpora are the ones that our system will find most relevant.

3.1 Corpora Creation

In our first step of determining how corpora relate, we conflated our annotations so that all continuation-of utterances would be given the communicative action of the utterance that it continues. With this restriction, we obtained all utterances which were labeled as *response-to* and all utterances that were labeled as *addressed-to*.

To create our separate corpora out of the original corpus of utterances we collect all n-grams of length less than or equal to four[1]. For each n-gram, if it appeared in an utterance that was labeled as an *addressed-to* we appended it to the *addressed-to* corpus. If the n-gram appeared in an utterance labled *response-to* we appended it to the *response-to* corpus. After this process we order these n-grams by the amount of times they are contained in an utterance with the given communicative action.

To determine the likelihood of a new utterance belonging to a given communicative action we iterate through every n-gram in that utterance. This results in a set of n-grams of the form (w_i) which are contained in the utterance.

$$utterance \rightarrow \{w_0, w_1, w_2, ..., w_m\}$$

[1]This heuristic reduces the time complexity of n-gram collection from a polynomial time problem to a linear one

3.2 Ratios and Differences

For every given n-gram, we compute two metrics:

Frequency Difference the frequency difference measures the distance between the frequencies of an n-gram in two corpora.

Frequency Ratio the frequency ratio measures the ratio of frequencies of an n-gram in two corpora.

These metrics serve as an indication of which corpora the n-gram is more likely to belong.

$$freq_x(w_i) = \text{the frequency of n-gram } w_i \text{ in corpora x}$$
$$freqDifference(w_i) = freq_a(w_i) \text{ - } freq_b(w_i)$$
$$freqRatio(w_i) = freq_a(w_i)/freq_b(w_i)$$

To compute the likelihood of an utterance to belong to one corpus rather than the other, we take the sum of the $freqDifference$s (or the product of the frequency ratios)of all the n-grams that appeared in both corpora.

$$commValence(utterance)=\sum_{i=0}^{m}(freqDifference(w_i))$$

or if ratios are being used

$$commValence(utterance)=\prod_{i=0}^{m}(freqRatio(w_i))$$

This method provided the basis of our computation, but it attributed no special weighting to the length of the n-gram. For this reason, we decided to create a constant we shall denote as $wtPow$. This constant was derived empirically and is intended to determine how relevant the size of the n-gram is in its effect on the communicative action valence. The results of this analysis can be found in Figures [1 & 1].

$$commValence(utterance) = \sum_{i=0}^{m}\{(freqDifference(w_i)) * length(w_i)^{wtPow}\}$$
$$commValence(utterance) = \prod_{i=0}^{m}\{(freqRatio(w_i)) * length(w_i)^{wtPow}\}$$

As expected, if the length of the n-gram were completely irrelevant the value of $wtPow$ would be zero and the equation would flatten down to that without a $wtPow$.

3.3 Multiple Corpora Comparison

To extend our system we first deflated our communicative actions to incorporate the *continuation-of* tag. This increases our total to three tags, but our earlier system has no support for more than two classifications. In order to accurately annotate an utterance when there are multple annotations available we modified our notion of $commValence$. For each classification group we create an associated valence.

$$freq_k(w_i) = \text{frequency of n-gram } w_i \text{ in class k}$$
$$valence_k(utterance) = \sum_{i=0}^{m} \{(freq_k(w_i)) * length(w_i)^{wtPow}\}$$

Alternatively, if we were to use frequency ratios the following would apply.

$$valence_k(utterance) = \prod_{i=0}^{m} \{(freq_k(w_i)) * length(w_i)^{wtPow}\}$$

In either case, to determine the valence of a particular utterance we need only to determine the valence with the highest value.

```
for(int k=0; k<CLASS_AMT; k++){
    if(valence(k,utterance) = maxValence){
        utterance.setClass(k);
    }
}
```

4 Results

Our preliminary results employing this method are promising. As an upperbound for our analysis we computed the observed agreement of our annotators across all of the communicative actions, 73.59%. Our F-Scores for *response-to* and *addressed-to* (as seen in Figure [1] approach this value as the value of $wtPow$ increases.

Table 1: F-Scores for All Automatic Additive Communicative Actions (Wt-Pow=10)

	03/29	04/26	04/28	04/30	05/03	05/04	05/06
Precision(R)	48.80%	53.51%	65.30%	51.55%	55.63%	68.82%	56.64%
Recall(R)	62.79%	55.45%	64.05%	58.08%	58.33%	50.66%	55.67%
F-Score(R)	54.92%	54.46%	57.94%	54.62%	56.95%	58.36%	56.15%
Precision(A)	52.75%	52.14%	52.90%	33.49%	53.85%	23.77%	30.77%
Recall(A)	50.66%	66.30%	38.24%	63.64%	64.17%	75.93%	67.29%
F-Score(A)	51.68%	58.37%	27.23%	43.89%	58.56%	36.21%	42.23%
Precision(C)	38.10%	36.67%	21.14%	38.46%	12.00%	25.37%	34.18%
Recall(C)	25.81%	21.57%	51.08%	14.71%	6.82%	15.60%	15.08%
F-Score(C)	30.77%	27.16%	45.42%	21.28%	8.70%	19.32%	20.93%

Table 2: F-Scores for Conflated Automatic Additive Communicative Actions (WtPow=10)

	03/29	04/26	04/28	04/30	05/03	05/04	05/06
Precision(R)	61.90%	69.72%	51.69%	79.62%	70.55%	89.82%	85.37%
Recall(R)	75.58%	69.09%	83.56%	73.36%	71.53%	65.17%	72.16%
F-Score(R)	68.06%	69.40%	63.87%	76.36%	71.04%	75.53%	78.21%
Precision(A)	62.94%	63.44%	77.22%	52.34%	65.25%	37.74%	46.71%
Recall(A)	47.14%	64.13%	41.64%	60.91%	64.17%	74.07%	66.36%
F-Score(A)	53.91%	63.78%	54.10%	56.30%	64.71%	50.00%	54.83%

From our experiments, we see that optimal value for $wtPow$ when using frequency differences is approximately 10, after which the F-Scores show no substantial improvement. This indicates that, all things being equal, a longer n-gram is more relevant than a shorter n-gram when determining which communicative action an utterance belongs to.

Similarly as $wtPow$ is increased for frequency ratios, there is only slow growth in F-Scores starting from a $wtPow$ of 0. For very large values of $wtPow$ both metrics give nearly the same values, because the long n-grams become all that is relevant.

The F-Score values for classification of *continuation-of* utterances are much lower than those of the other two classes. This is due in part because a *continuation-of* utterance can function as any of the three classes. From

Table 3: F-Scores for All Automatic Multiplicative Communicative Actions (WtPow=0)

	03/29	04/26	04/28	04/30	05/03	05/04	05/06
Precision(R)	46.44%	48.94%	39.81%	49.67%	55.93%	68.18%	56.05%
Recall(R)	73.26%	62.73%	76.71%	66.38%	68.75%	59.37%	65.29%
F-Score(R)	56.85%	54.98%	52.42%	56.82%	61.68%	63.47%	60.32%
Precision(A)	57.24%	50.00%	71.43%	32.60%	56.10%	24.68%	28.64%
Recall(A)	38.33%	54.35%	35.84%	53.64%	57.50%	70.37%	53.27%
F-Score(A)	45.91%	52.08%	47.73%	40.55%	56.79%	36.54%	37.25%
Precision(C)	22.67%	25.00%	26.83%	44.44%	10.53%	32.08%	34.43%
Recall(C)	13.71%	9.80%	17.89%	11.76%	4.55%	15.60%	11.73%
F-Score(C)	17.09%	14.08%	21.47%	18.60%	6.35%	20.99%	17.50%

Table 4: F-Scores for Conflated Automatic Multiplicative Communicative Actions (WtPow=0)

	03/29	04/26	04/28	04/30	05/03	05/04	05/06
Precision(R)	62.88%	64.35%	51.10%	77.46%	70.13%	88.97%	82.42%
Recall(R)	79.46%	67.27%	84.47%	72.05%	75.00%	61.74%	72.51%
F-Score(R)	70.20%	65.78%	63.68%	74.66%	72.48%	72.90%	77.15%
Precision(A)	66.67%	58.62%	77.33%	49.21%	67.27%	35.27%	43.66%
Recall(A)	46.70%	55.43%	39.59%	56.36%	61.67%	73.15%	57.94%
F-Score(A)	54.93%	56.98%	52.37%	52.54%	64.35%	47.59%	49.80%

a simplistic standpoint, an n-gram that appeared in a *continuation-of* utterance could have also appeared in a combined utterance where the speaker never separated his utterance into two turns.

5 Problems and Improvement

This method of corpora comparison is heavily dependent on the quality of the annotators. Due to the fact that the system relies on the classification of the data by the annotators, we can only hope to get values as good as the agreement of the annotators will allow.

This method may be aided by the addition of a small set of rules which factor into the valence of an utterance. This would allow us to create a

statistical system that is aided by our understood knowledge of how communicative acts should be used.

Another problem that this method ignores is that two corpora may differ in size greatly. A real world example would be comparing the "corpus" of a small amount utterances to the rest of the discussion. This set up would allow an analyst to determine if there is a term in a small range that is more relevant than anywhere else. This may correspond to the notion of "burstiness"[KNRT05].

6 Other Applications

Corpora comparison by n-gram frequency may have many uses in the future. This technique can be used in a variety of problems in natural language processing from determining the topic of a text to resolving the author of a portion of work.

To determine topic from a sample corpus one method could compare the sample text to a known general corpus. The words that appear much more frequently in the sample corpus than in the general corpus are the ones that are most indicative of the topic of the sample text. Moreover, if some of these terms appear more often than even the predefined stopwords do, there is a much higher indication of the word being related to the topic of conversation.

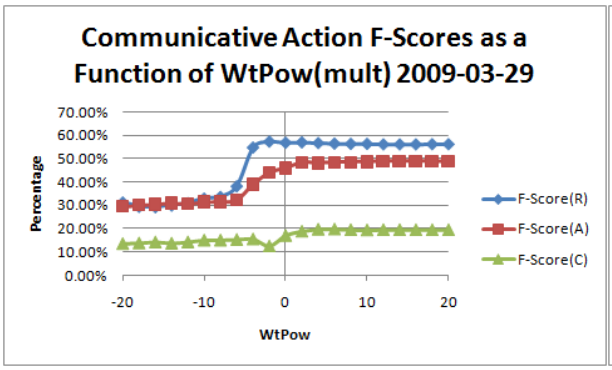
References

[JSE+05] Guoqian Jiang, Hitomi Sato, Akira Endoh, Katsuhiko Ogasawara, and Tsunetaro Sakurai. *Extraction of Specific Nursing Terms Using Corpora Comparison.* 2005.

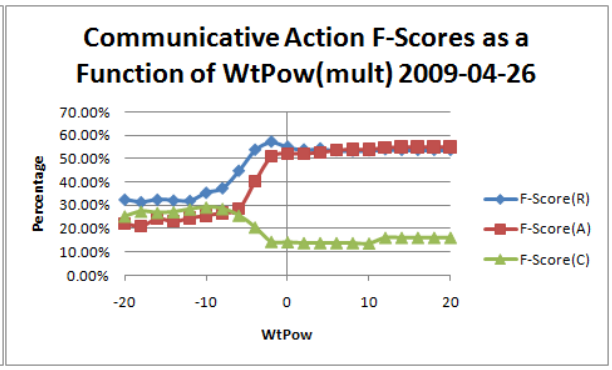
[KNRT05] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. *On the Bursty Evolution of Blogspace. World Wide Web*, 8(2):159–178, June 2005.

Figure 1: Experimental Results of Varying the Weight of the N-Gram Length on the F-Scores for All Communicative Actions: *response-to*, *addressed-to*, and *continuation-of* using Frequency Ratios

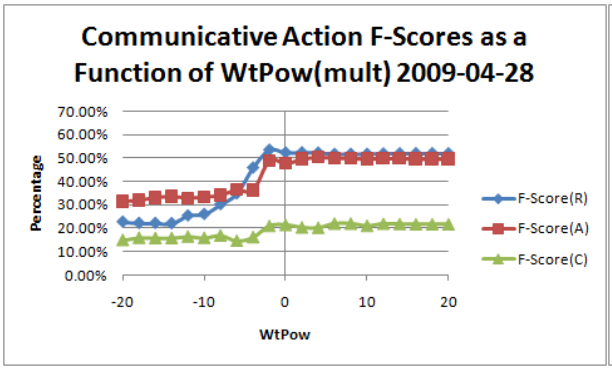
(a) 03/29/09



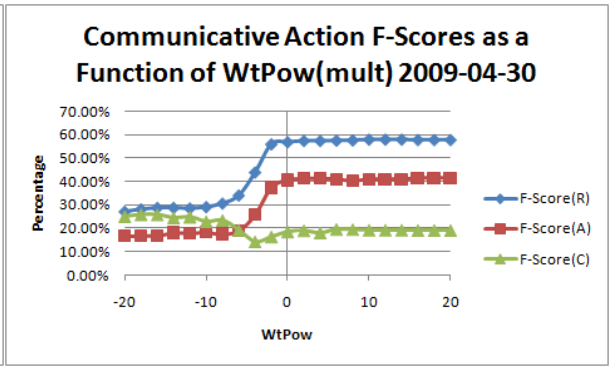
(b) 04/26/09



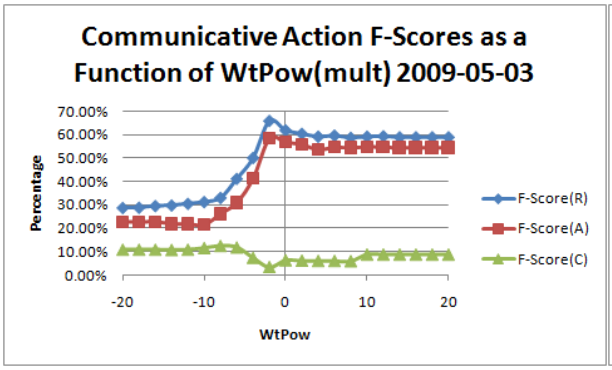
(c) 04/28/09



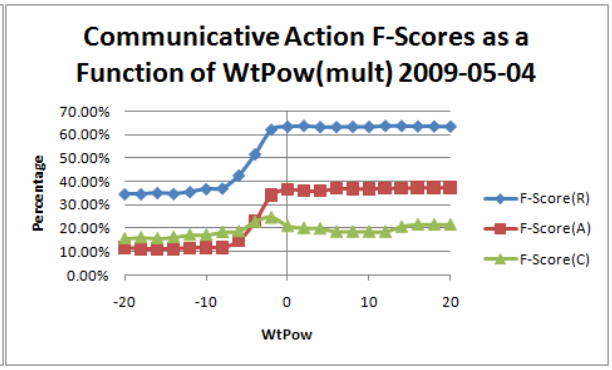
(d) 04/30/09

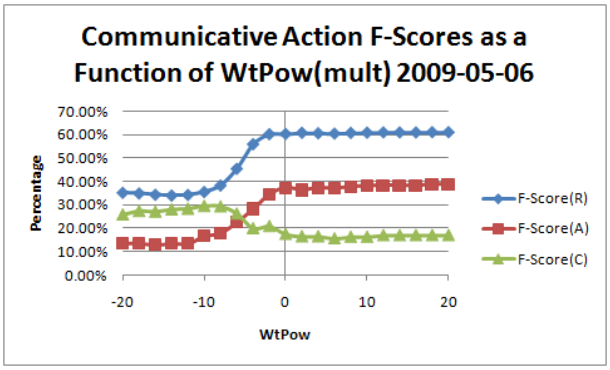


(e) 05/03/09



(f) 05/04/09

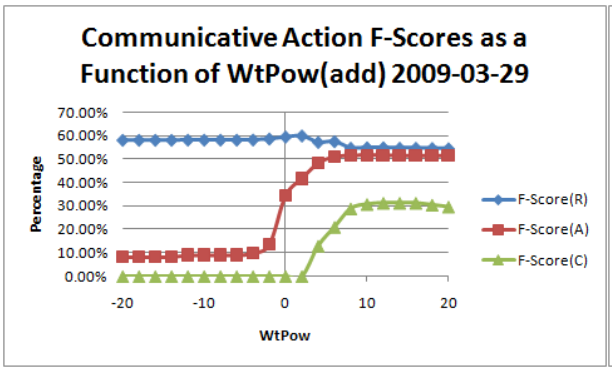



Communicative Action F-Scores as a Function of WtPow(mult) 2009-05-06
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Percentage
-20
-10
0
10
20
WtPow
F-Score(R)
F-Score(A)
F-Score(C)

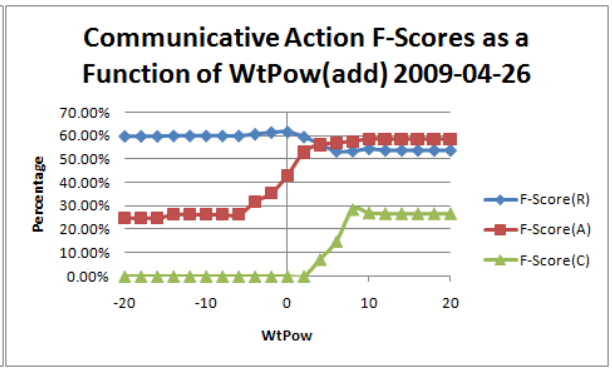
(g) 05/06/09

Figure 1: Experimental Results of Varying the Weight of the N-Gram Length on the F-Scores for All Communicative Actions: *response-to*, *addressed-to*, and *continuation-of* using Frequency Differences

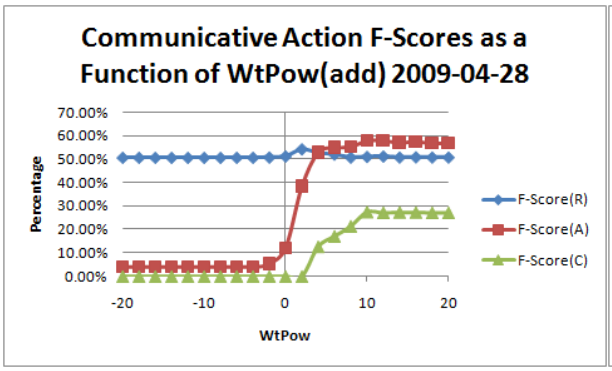
(a) 03/29/09



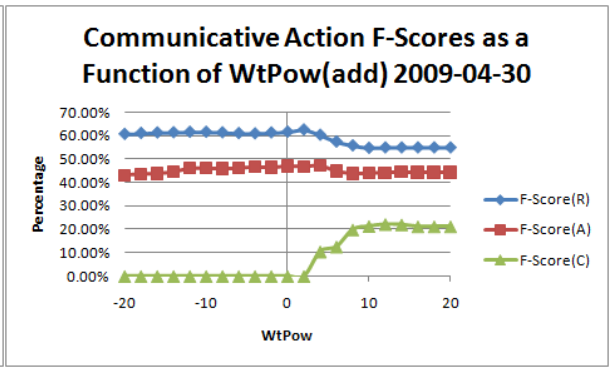
(b) 04/26/09



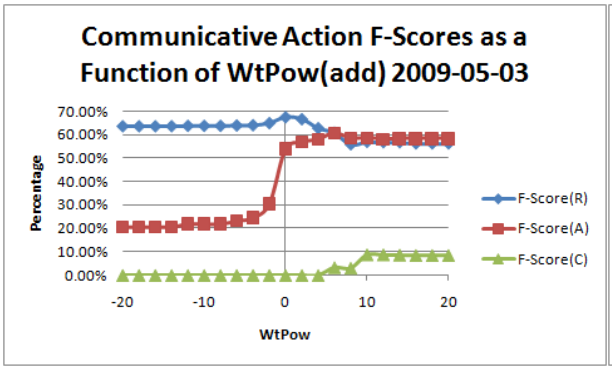
(c) 04/28/09



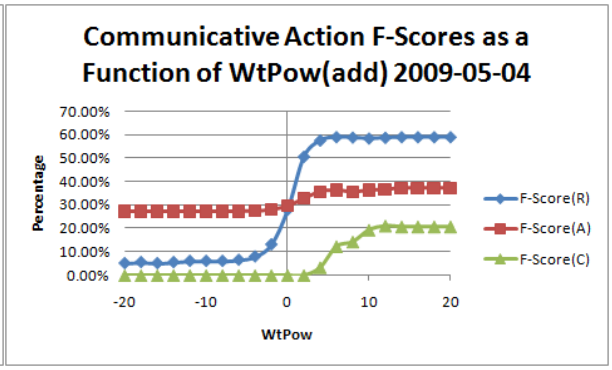
(d) 04/30/09

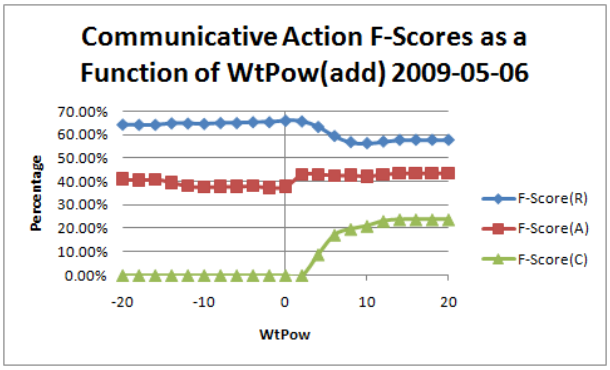


(e) 05/03/09



(f) 05/04/09

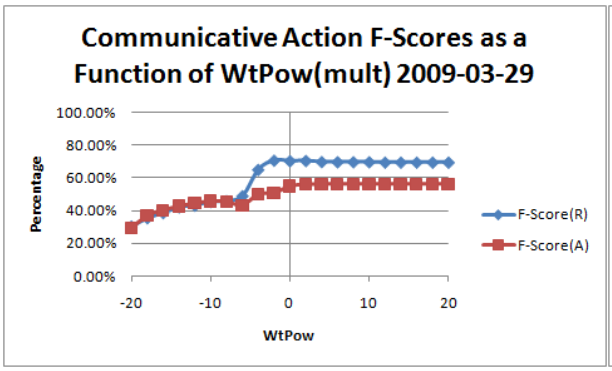



Communicative Action F-Scores as a Function of WtPow(add) 2009-05-06
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Percentage
-20
-10
0
10
20
WtPow
F-Score(R)
F-Score(A)
F-Score(C)

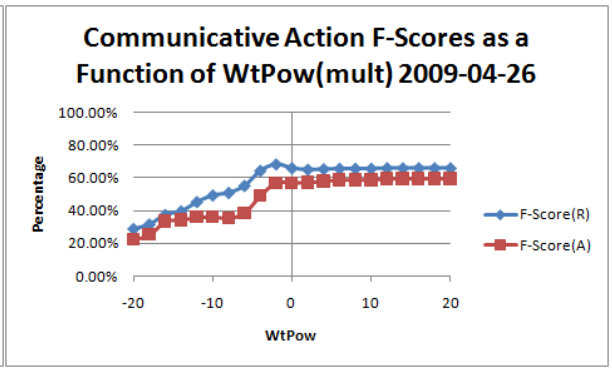
(g) 05/06/09

Figure 1: Experimental Results of Varying the Weight of the N-Gram Length on the F-Scores for Conflated Communicative Actions: *response-to* and *addressed-to* using Frequency Ratios

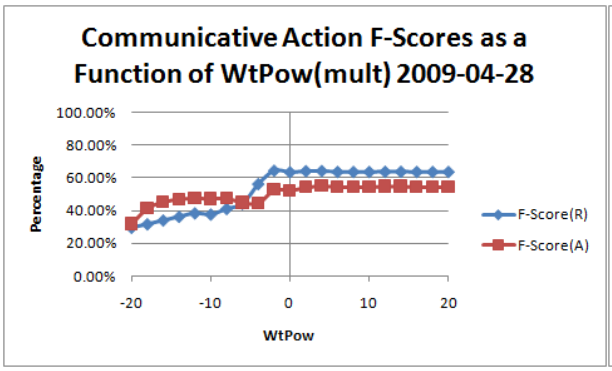
(a) 03/29/09



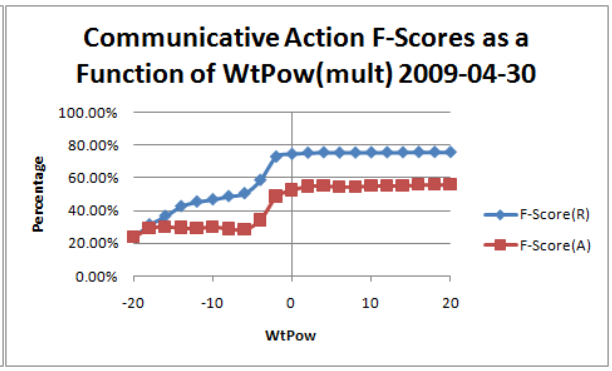
(b) 04/26/09



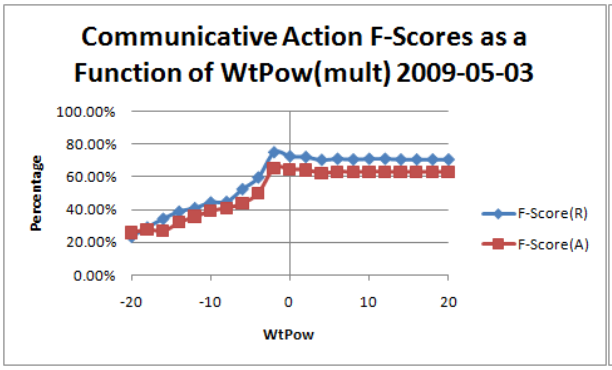
(c) 04/28/09



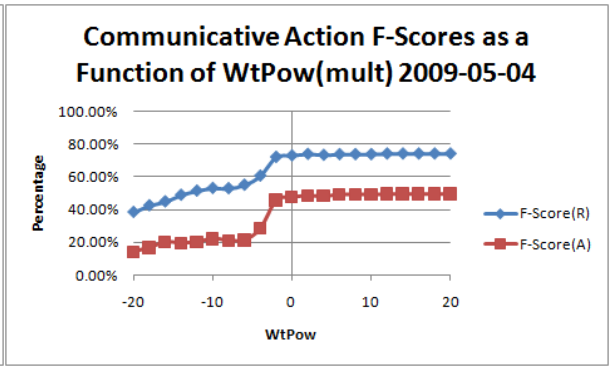
(d) 04/30/09

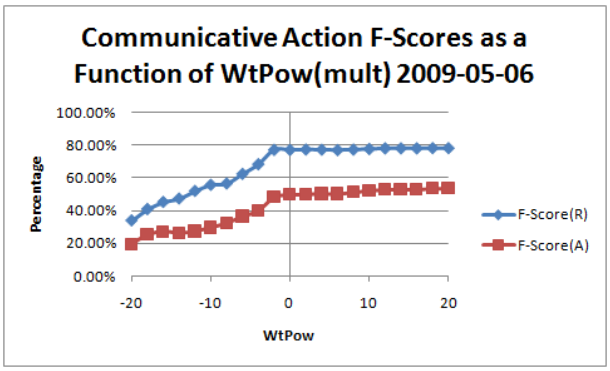


(e) 05/03/09



(f) 05/04/09

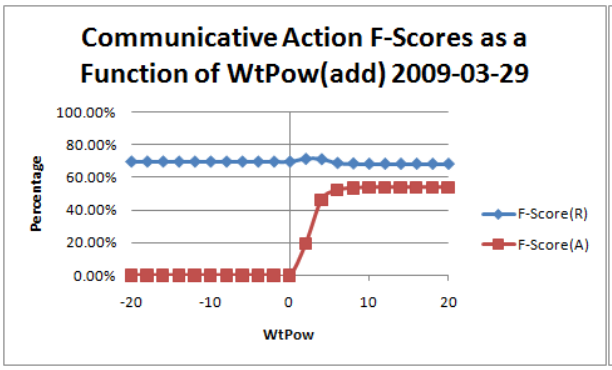



Communicative Action F-Scores as a Function of WtPow(mult) 2009-05-06
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
Percentage
-20
-10
0
10
20
WtPow
F-Score(R)
F-Score(A)

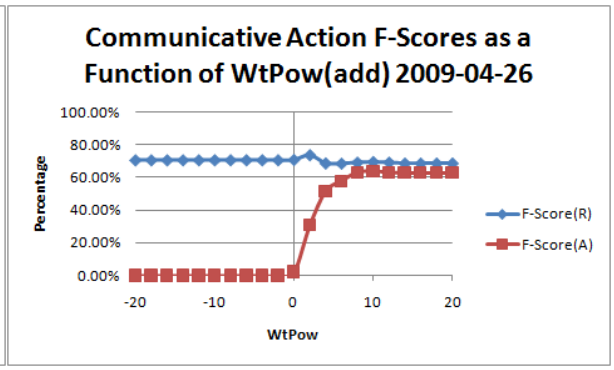
(g) 05/06/09

Figure 1: Experimental Results of Varying the Weight of the N-Gram Length on the F-Scores for Conflated Communicative Actions: *response-to* and *addressed-to* using Frequency Differences

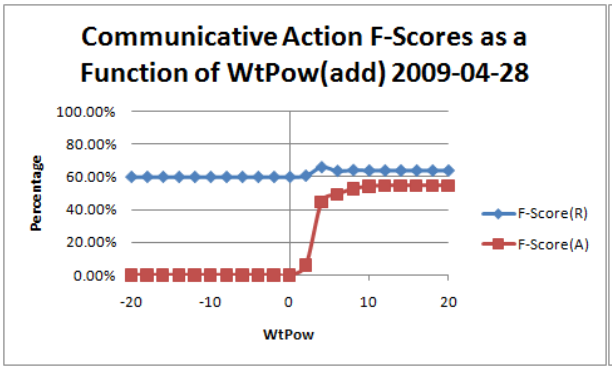
(a) 03/29/09



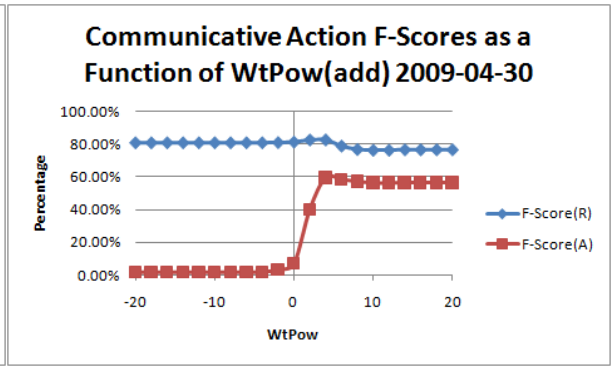
(b) 04/26/09



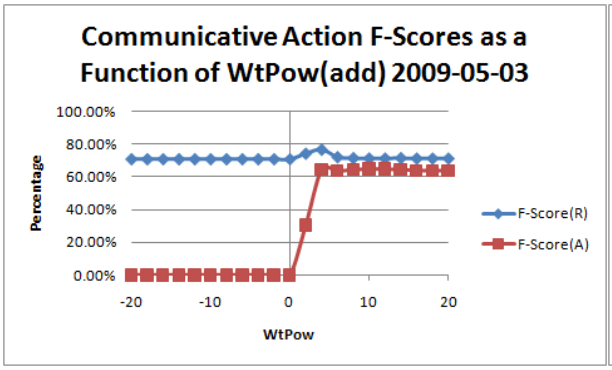
(c) 04/28/09



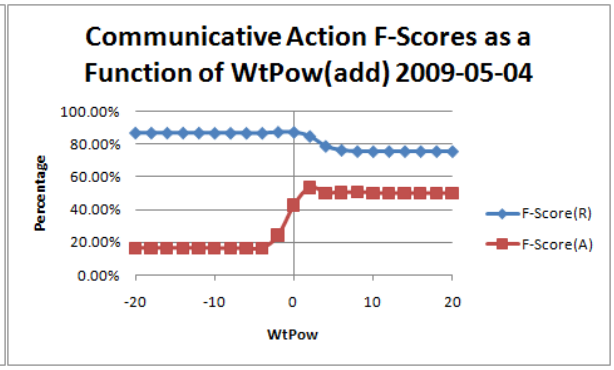
(d) 04/30/09

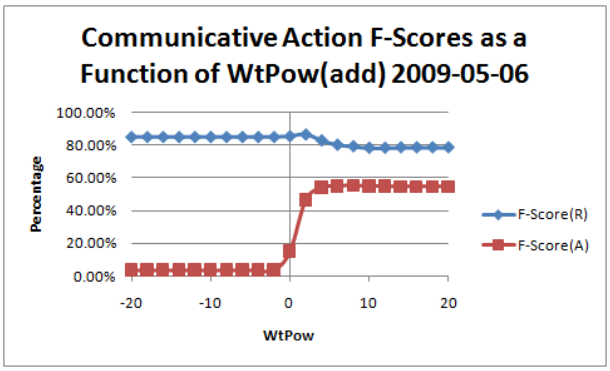


(e) 05/03/09



(f) 05/04/09




Communicative Action F-Scores as a Function of WtPow(add) 2009-05-06
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
Percentage
-20
-10
0
10
20
WtPow
F-Score(R)
F-Score(A)

(g) 05/06/09