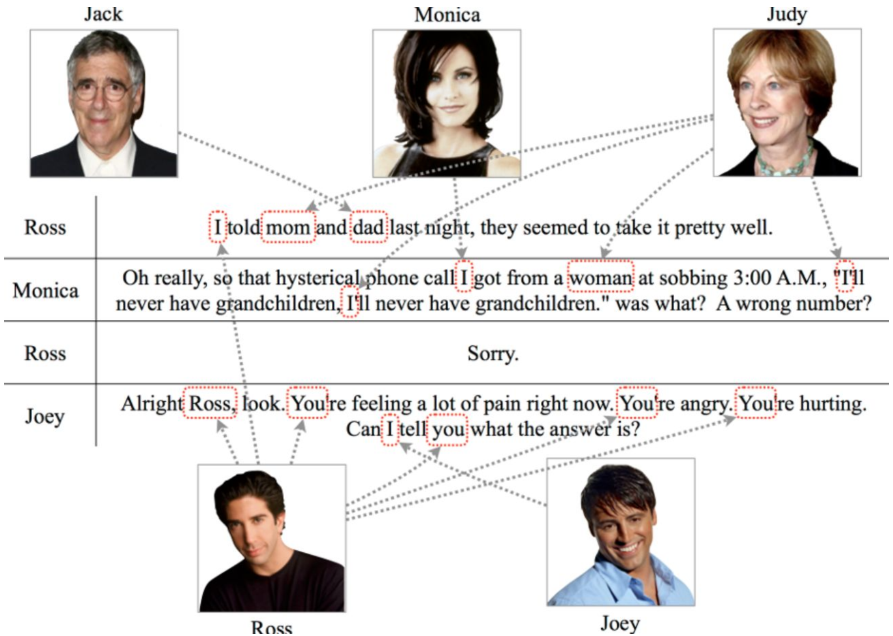
Semeval Task 4

Character Identification on Multiparty Dialogues

Casey Beaird, Chase Greco, Brandon Watts

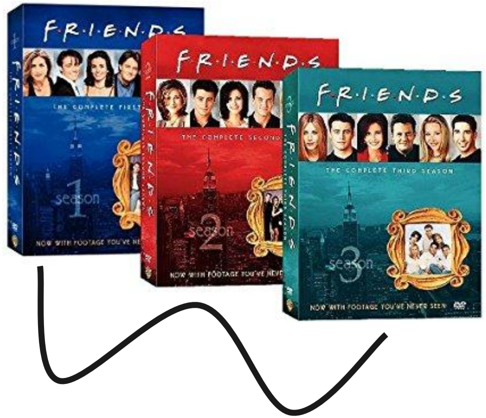
Problem Statement

Assign each mention in the script of the popular TV show *Friends* to its entity, who may or may not participate in the dialogue.



Why it's difficult..

- It is cross document (multiple seasons)
- There are a TON of mentions in a tv script
- The data is absolutely *horrible* (more on that later)



Series Cast complete, awaiting verification

Jennifer Aniston	...	Rachel Green (236 epi
Courteney Cox	...	Monica Geller / ... (23
Lisa Kudrow	...	Phoebe Buffay / ... (23
Matt LeBlanc	...	Joey Tribbiani (236 ep
Matthew Perry	...	Chandler Bing (236 ep
David Schwimmer	...	Dr. Ross Geller / ... (2
James Michael Tyler	...	Gunther (160 episodes
Elliott Gould	...	Jack Geller (21 episod
Christina Pickles	...	Judy Geller (20 episod
Maggie Wheeler	...	Janice (19 episodes, 1
Paul Rudd	...	Mike Hannigan (18 epi
Jane Sibbett	...	Carol Willick (15 episo
Helen Baxendale	...	Emily Waltham (14 ep

Jessica Hecht	...	Susan Bunch (12 episodes, 1994-200
June Gable	...	Estelle Leonard / ... (11 episodes, 19
Tom Selleck	...	Dr. Richard Burke (10 episodes, 1996
Aisha Tyler	...	Charlie Wheeler (9 episodes, 2003)
Giovanni Ribisi	...	Frank Buffay Jr. / ... (9 episodes, 199
Lauren Tom	...	Julie (8 episodes, 1995-2002)
Eddie Cahill	...	Tag Jones (7 episodes, 2000-2001)
Bonnie Somerville	...	Mona (7 episodes, 2001-2002)
Cole Sprouse	...	Ben Geller (8 episodes, 2000-2002)
Jon Favreau	...	Pete Becker (6 episodes, 1997)
Amanda Carlin	...	Dr. Long (6 episodes, 2001-2002)
Steven Eckholdt	...	Mark Robinson (6 episodes, 1997-20
Paget Brewster	...	Kathy (6 episodes, 1997-1998)
Mitchell Whitfield	...	Barry Farber (6 episodes, 1994-2000

Debra Jo Rupp	...	Alice Knight Buffay / ... (6 episodes, 1997-1998)
Steve Ireland	...	Mr. Zelner (6 episodes, 1999-2004)
Hank Azaria	...	David (5 episodes, 1994-2003)
Tate Donovan	...	Joshua Burgin (5 episodes, 1998)
Elle Macpherson	...	Janine Lecroix (6 episodes, 1999-2002)
Anna Faris	...	Erica (5 episodes, 2004)
Alexandra Holden	...	Elizabeth Stevens (5 episodes, 2000)
Larry Hankin	...	Mr. Heckles (5 episodes, 1994-1996)
Mike Hagerty	...	Mr. Treeger (5 episodes, 1995-2001)
Morgan Fairchild	...	Nora Tyler Bing (5 episodes, 1995-2001)
Charles Thomas Allen	...	Ben Geller / ... (5 episodes, 1996-1999)
John Christopher Allen	...	Ben Geller / ... (5 episodes, 1996-1999)
Cynthia Mann	...	Jasmine (5 episodes, 1994-1998)

Methods

We started by establishing a simple baseline by choosing a speakers most likely tag given a word. From the papers we read, this method seemed to be widely regarded as the method of choice, so we thought it would make a nice baseline for our machine learning algorithms.

We used information from the mentions to create custom feature vectors incorporating both lexical and orthographic properties. We tested a variety of machine learning algorithms in WEKA including Naïve Bayes, SVM, and C.45.

Most Likely Tag Baseline

The most likely tag baseline selects the tag (t_i hat) by selecting the tag (t_i) with the highest probability given the joint probability of the speaker (s) and word (w). To create this baseline we had to create a probability matrix to determine $P(t_i|s,w)$. This was created by adding all the time a speaker mentioned a specific entity.

$$\hat{t_i} = argmax_{t_i}(P(t_i|s,w))$$

Machine Learning

The first step in the machine learning phase was to somehow convert these mentions into a feature vector. Then we could feed these features to a variety of machine learning algorithms and see how they perform. We started by trying to build our own word encodings using one-hot vectors and a skip-gram model with a window size of 2. This caused our feature vectors to be completely **massive** and **extremely sparse**. This would have taken ages to train so we ended up utilizing word2vec which was more efficient.

```
/friends-s02e07   0   0     Wha   NNP        (TOP(NP*    Wha    -   -   Ross_Geller
/friends-s02e07   0   1     ...   :               *))    ...    -   -   Ross_Geller

/friends-s02e07   0   0     you   PRP  (TOP(S(S(NP*)    you    -   -   Ross_Geller
/friends-s02e07   0   1     're   VBP            (VP*    be     -   -   Ross_Geller
/friends-s02e07   0   2     uh    UH        (S(INTJ*)    uh     -   -   Ross_Geller
/friends-s02e07   0   3     ,     ,                 *    ,      -   -   Ross_Geller
/friends-s02e07   0   4     you   PRP           (NP*)    you    -   -   Ross_Geller
/friends-s02e07   0   5     're   VBP         (VP*)))    be     -   -   Ross_Geller
/friends-s02e07   0   6     ,     ,                 *    ,      -   -   Ross_Geller
/friends-s02e07   0   7     you   PRP           (NP*)    you    -   -   Ross_Geller
/friends-s02e07   0   8     're   VBP            (VP*    be     -   -   Ross_Geller
/friends-s02e07   0   9     over  IN             (PP*    over   -   -   Ross_Geller
/friends-s02e07   0   10    me    PRP         (NP*)))    I      -   -   Ross_Geller
/friends-s02e07   0   11    ?     .               *))    ?      -   -   Ross_Geller
```

$$v = [k_i, j_i, s_i, w_i, e_i]$$

Feature Vectors

The feature vectors are described as:

$$v = [k_i, j_i, s_i, w_i, e_i]$$

Where
k_i = Season
j_i = Episode
s_i = Speaker
w_i = word2vec representation of the mention
e_i = Entity

The idea was to combine both lexical and orthographic information into a single vector

Machine Learning Algorithms

Naïve Bayes	SVM	C.45 (Decision Tree)
Typically a poor performer, but often cited as well suited to NLP tasks	Cited in the literature as the highest performing algorithm for this task	Cited in the literature as another useful algorithm when performing this task

Data

One of the biggest challenges was the data. There were many of discrepancies that caused (and continues to cause) issues.

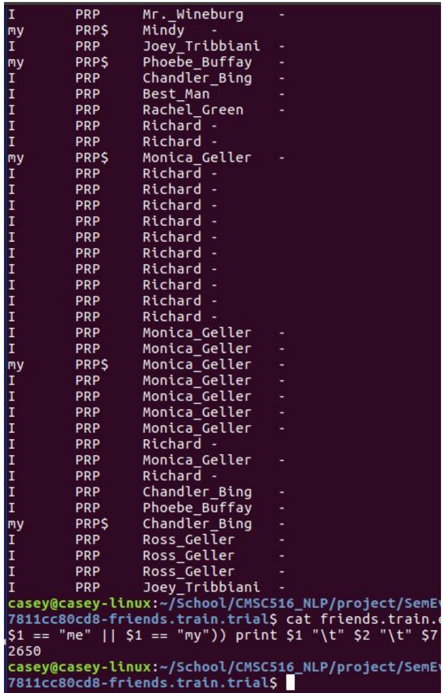
```
/friends-s02e07  0  0   Ohhhhhhhh  NNP           (TOP(NP*   Ohhhhhhhh  -  -  Rachel_Green  (MISC*  (335)
/friends-s02e07  0  1         God  NNP                  *         God  -  -  Rachel_Green      *)      -
/friends-s02e07  0  2           .    .                *))           .  -  -  Rachel_Green       *  (335)

/friends-s02e07  0  0         Wha  NNP           (TOP(NP*         Wha  -  -  Ross_Geller        *      -
/friends-s02e07  0  1         ...    :                *))         ...  -  -  Ross_Geller        *      -

/friends-s02e07  0  0         you  PRP     (TOP(S(S(NP*)        you  -  -  Ross_Geller        *      -
/friends-s02e07  0  1         're  VBP               (VP*          be  -  -  Ross_Geller        *      -
/friends-s02e07  0  2          uh   UH          (S(INTJ*)          uh  -  -  Ross_Geller        *      -
/friends-s02e07  0  3           ,    ,                  *           ,  -  -  Ross_Geller        *  (335)
/friends-s02e07  0  4         you  PRP              (NP*)         you  -  -  Ross_Geller        *      -
/friends-s02e07  0  5         're  VBP           (VP*))))          be  -  -  Ross_Geller        *      -
/friends-s02e07  0  6           ,    ,                  *           ,  -  -  Ross_Geller        *  (306)
/friends-s02e07  0  7         you  PRP              (NP*)         you  -  -  Ross_Geller        *      -
/friends-s02e07  0  8         're  VBP               (VP*          be  -  -  Ross_Geller        *      -
/friends-s02e07  0  9        over   IN               (PP*        over  -  -  Ross_Geller        *  (335)
/friends-s02e07  0  10         me  PRP            (NP*)))           I  -  -  Ross_Geller        *      -
/friends-s02e07  0  11          ?    .                *))           ?  -  -  Ross_Geller        *      -

/friends-s02e07  0  0         Ohh   RB  (TOP(FRAG(ADVP*)          ohh  -  -  Rachel_Green       *      -
/friends-s02e07  0  1           ,    ,                  *           ,  -  -  Rachel_Green       *      -
/friends-s02e07  0  2         ohh   NN              (NP*)         ohh  -  -  Rachel_Green       *      -
/friends-s02e07  0  3           .    .                *))           .  -  -  Rachel_Green       *      -
```

No for real we mean Data Issues

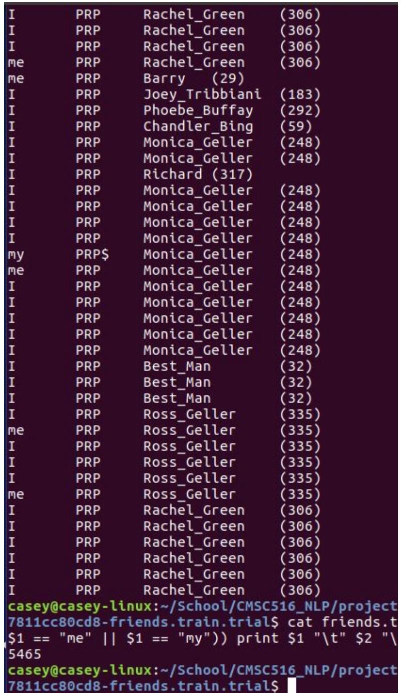
8461	,	,	Ross_Geller	(335)
8462	,	,	Ross_Geller	(306)
8467	.	.	Ross_Geller	(335)
8468	,	,	Ross_Geller	(306)
8469	.	.	Ross_Geller	(335)
8475	've	VBP	Ross_Geller	(306)
8477	had	VBD	Rachel_Green	(335)
8478	for	IN	Rachel_Green	(306)
8479	.	.	Ross_Geller	(335)
8483	.	.	Rachel_Green	(335)
8484	.	.	Ross_Geller	(335)
8487	...	:	Ross_Geller	(306)
8492	.	.	Rachel_Green	(59)
8493	.	.	Ross_Geller	(59)
8495	,	,	Ross_Geller	(197)
8496	need	VBP	Ross_Geller	(197)
8498	,	,	Ross_Geller	(197)
8499	.	.	Ross_Geller	(335)
8500	'm	VBP	Ross_Geller	(306)
8502	and	CC	Ross_Geller	(335)
8503	so	IN	Ross_Geller	(335)
8507	.	.	Ross_Geller	(306)
8509	.	.	Ross_Geller	(335)
8511	got	VBD	Julie	(197)
8512	.	.	Julie	(197)
8513	'll	MD	Ross_Geller	(197)
8515	,	,	Rachel_Green	(335)
8516	,	,	Rachel_Green	(197)
8522	've	VBP	Ross_Geller	(335)
8524	'm	VBP	Ross_Geller	(335)
8826	minute	NN	Monica_Geller	(292
9467	Sales	NNS	Phoebe_Buffay	349)
9691	zoo	NN	Lipson	(212
10117	and	CC	Joey_Tribbiani	(59)
10122	is	VBZ	Chandler_Bing	(183)

```
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Ross_Geller	(335)
I	PRP	Ross_Geller	(335)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Ross_Geller	(335)
I	PRP	Rachel_Green	(306)
I	PRP	Mr._Wineburg	(266)
I	PRP	Mindy	(242)
I	PRP	Rachel_Green	(306)
I	PRP	Mindy	(242)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Joey_Tribbiani	(183)
I	PRP	Phoebe_Buffay	(292)
I	PRP	Chandler_Bing	(59)
I	PRP	Monica_Geller	(248)
I	PRP	Monica_Geller	(248)
I	PRP	Richard	(317)
I	PRP	Monica_Geller	(248)
I	PRP	Monica_Geller	(248)
I	PRP	Monica_Geller	(248)
I	PRP	Monica_Geller	(248)
I	PRP	Monica_Geller	(248)
I	PRP	Monica_Geller	(248)
I	PRP	Monica_Geller	(248)
I	PRP	Monica_Geller	(248)
I	PRP	Monica_Geller	(248)
I	PRP	Best_Man	(32)
I	PRP	Best_Man	(32)
I	PRP	Best_Man	(32)
I	PRP	Ross_Geller	(335)
I	PRP	Ross_Geller	(335)
I	PRP	Ross_Geller	(335)
I	PRP	Ross_Geller	(335)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
I	PRP	Rachel_Green	(306)
casey@casey-linux:~/School/CMSC516_NLP/project/SemEval_Character-Identification-on-Multiparty-Dialogues/datasets-None
ll | tr -s ' ' | cut -d ' ' -f 4-12 | awk '{if ($9 != "-" && $1 == "I") print $1 "\t" $2 "\t" $7 "\t" $9}' |wc -l
4116
casey@casey-linux:~/School/CMSC516_NLP/project/SemEval_Character-Identification-on-Multiparty-Dialogues/datasets-None
```

Data issues and then issues with data?



```
I       PRP     Mr._Wineburg    -
my      PRP$    Mindy   -
I       PRP     Joey_Tribbiani  -
my      PRP$    Phoebe_Buffay   -
I       PRP     Chandler_Bing   -
I       PRP     Best_Man        -
I       PRP     Rachel_Green    -
I       PRP     Richard -
I       PRP     Richard -
my      PRP$    Monica_Geller   -
I       PRP     Richard -
I       PRP     Richard -
I       PRP     Richard -
I       PRP     Richard -
I       PRP     Richard -
I       PRP     Richard -
I       PRP     Richard -
I       PRP     Richard -
I       PRP     Richard -
I       PRP     Richard -
I       PRP     Monica_Geller   -
I       PRP     Monica_Geller   -
my      PRP$    Monica_Geller   -
I       PRP     Monica_Geller   -
I       PRP     Monica_Geller   -
I       PRP     Monica_Geller   -
I       PRP     Monica_Geller   -
I       PRP     Richard -
I       PRP     Monica_Geller   -
I       PRP     Richard -
I       PRP     Chandler_Bing   -
I       PRP     Phoebe_Buffay   -
my      PRP$    Chandler_Bing   -
I       PRP     Ross_Geller     -
I       PRP     Ross_Geller     -
I       PRP     Ross_Geller     -
I       PRP     Joey_Tribbiani  -
casey@casey-linux:~/School/CMSC516_NLP/project/SemE
7811cc80cd8-friends.train.trial$ cat friends.train.
$1 == "me" || $1 == "my")) print $1 "\t" $2 "\t" $7
2650
casey@casey-linux:~/School/CMSC516_NLP/project/SemE
7811cc80cd8-friends.train.trial$
```



```
I       PRP     Rachel_Green    (306)
I       PRP     Rachel_Green    (306)
I       PRP     Rachel_Green    (306)
me      PRP     Rachel_Green    (306)
me      PRP     Barry   (29)
I       PRP     Joey_Tribbiani  (183)
I       PRP     Phoebe_Buffay   (292)
I       PRP     Chandler_Bing   (59)
I       PRP     Monica_Geller   (248)
I       PRP     Monica_Geller   (248)
I       PRP     Richard (317)
I       PRP     Monica_Geller   (248)
I       PRP     Monica_Geller   (248)
I       PRP     Monica_Geller   (248)
I       PRP     Monica_Geller   (248)
my      PRP$    Monica_Geller   (248)
me      PRP     Monica_Geller   (248)
I       PRP     Monica_Geller   (248)
I       PRP     Monica_Geller   (248)
I       PRP     Monica_Geller   (248)
I       PRP     Monica_Geller   (248)
I       PRP     Monica_Geller   (248)
I       PRP     Best_Man        (32)
I       PRP     Best_Man        (32)
I       PRP     Best_Man        (32)
I       PRP     Ross_Geller     (335)
me      PRP     Ross_Geller     (335)
I       PRP     Ross_Geller     (335)
I       PRP     Ross_Geller     (335)
I       PRP     Ross_Geller     (335)
me      PRP     Ross_Geller     (335)
I       PRP     Rachel_Green    (306)
I       PRP     Rachel_Green    (306)
I       PRP     Rachel_Green    (306)
I       PRP     Rachel_Green    (306)
I       PRP     Rachel_Green    (306)
I       PRP     Rachel_Green    (306)
casey@casey-linux:~/School/CMSC516_NLP/project
7811cc80cd8-friends.train.trial$ cat friends.t
$1 == "me" || $1 == "my")) print $1 "\t" $2 "\
5465
casey@casey-linux:~/School/CMSC516_NLP/project
7811cc80cd8-friends.train.trial$
```

Issues with Data

- The conll format that was chosen does not have many usable parsers. A standardized format such as XML or JSON would have been much better to use.
- The data was not formatted correctly (tab delimited) which causes issues when we attempt to parse it
- Since the data is computer generated there are things it tags as entities which makes no logical sense (punctuation, interjections, etc.)
- The “training data” and the “test data” are exactly the same file. We did 10-fold cross validation to attempt to reduce sampling bias

Results

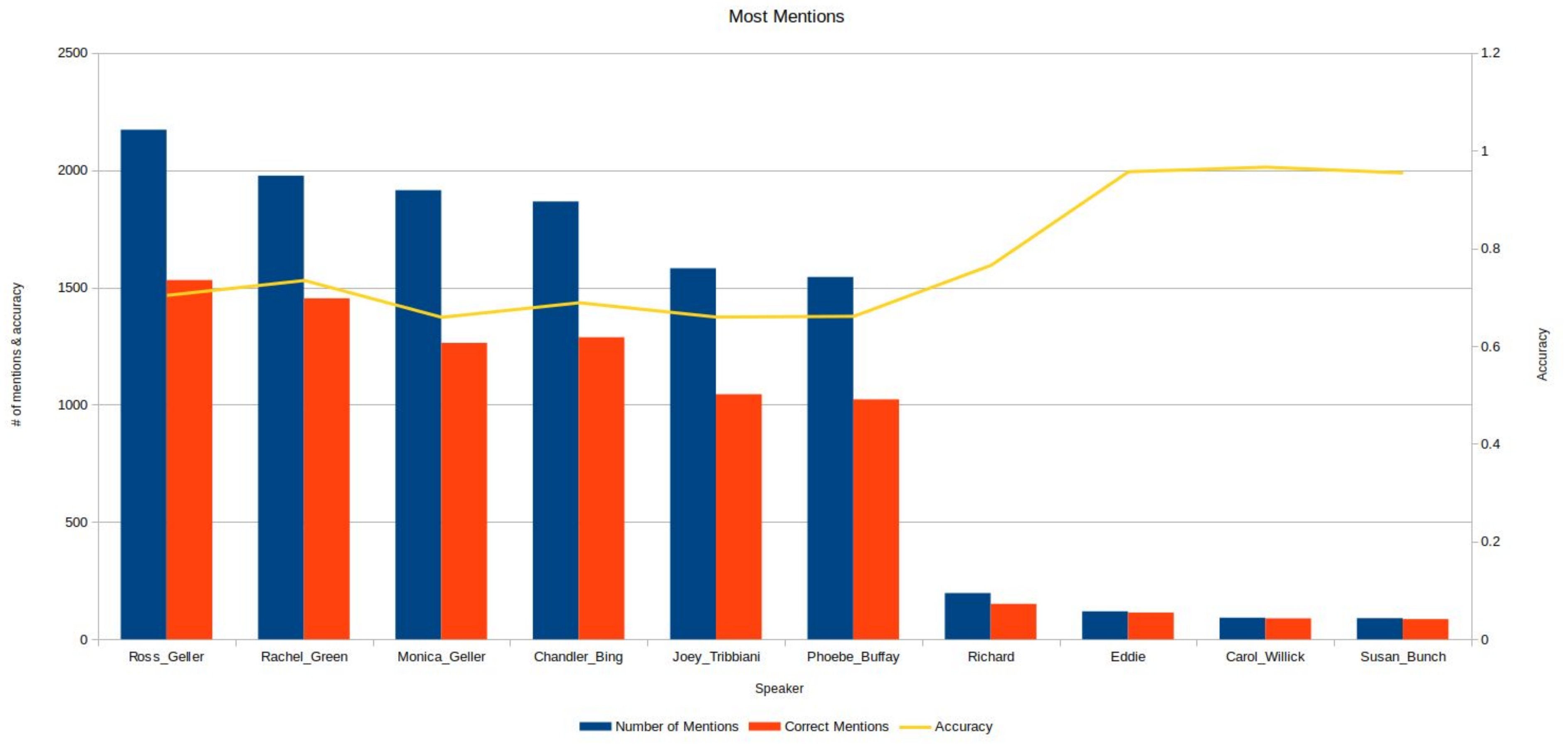
Most Likely Tag Baseline

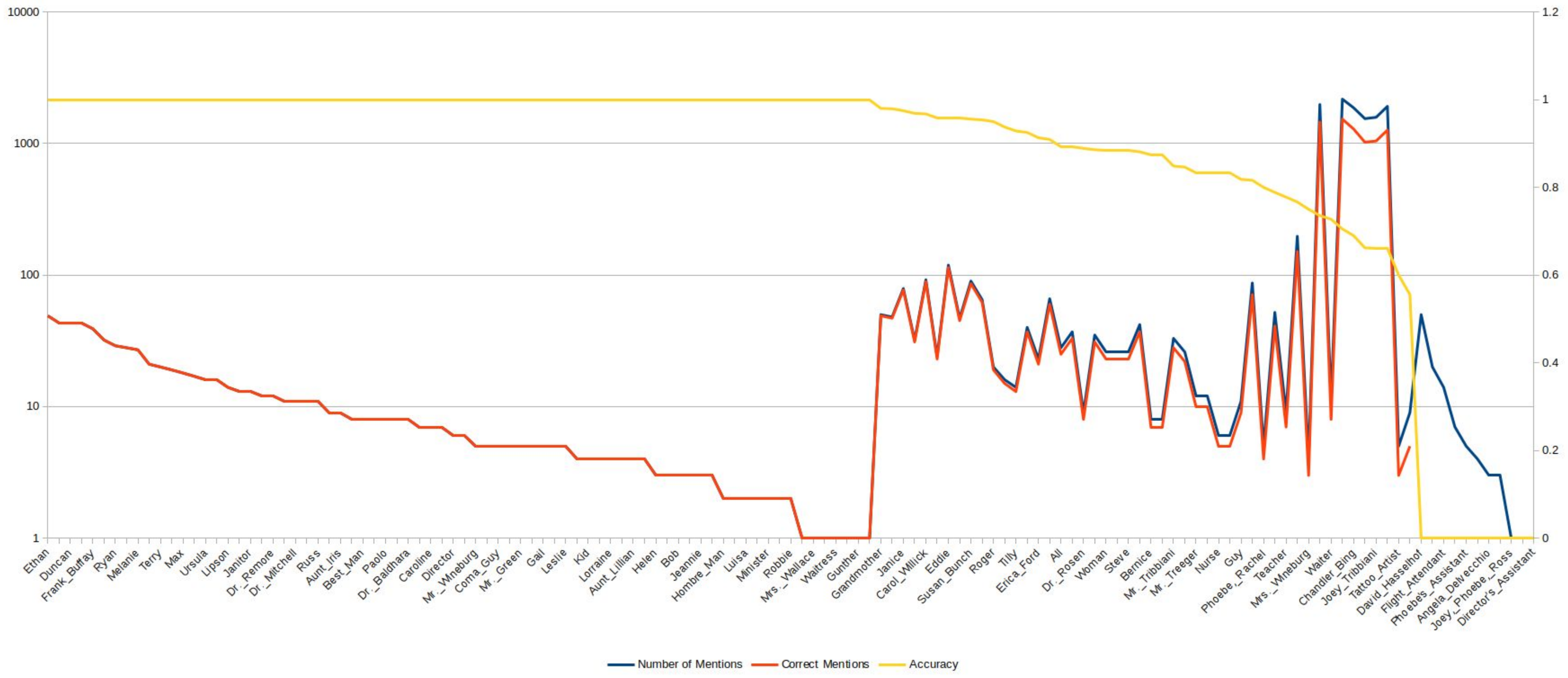
The most likely tag baseline actually performed relatively well...

- Word not found for speaker: 8
- Speaker not found: 109
- Other Errors: 12

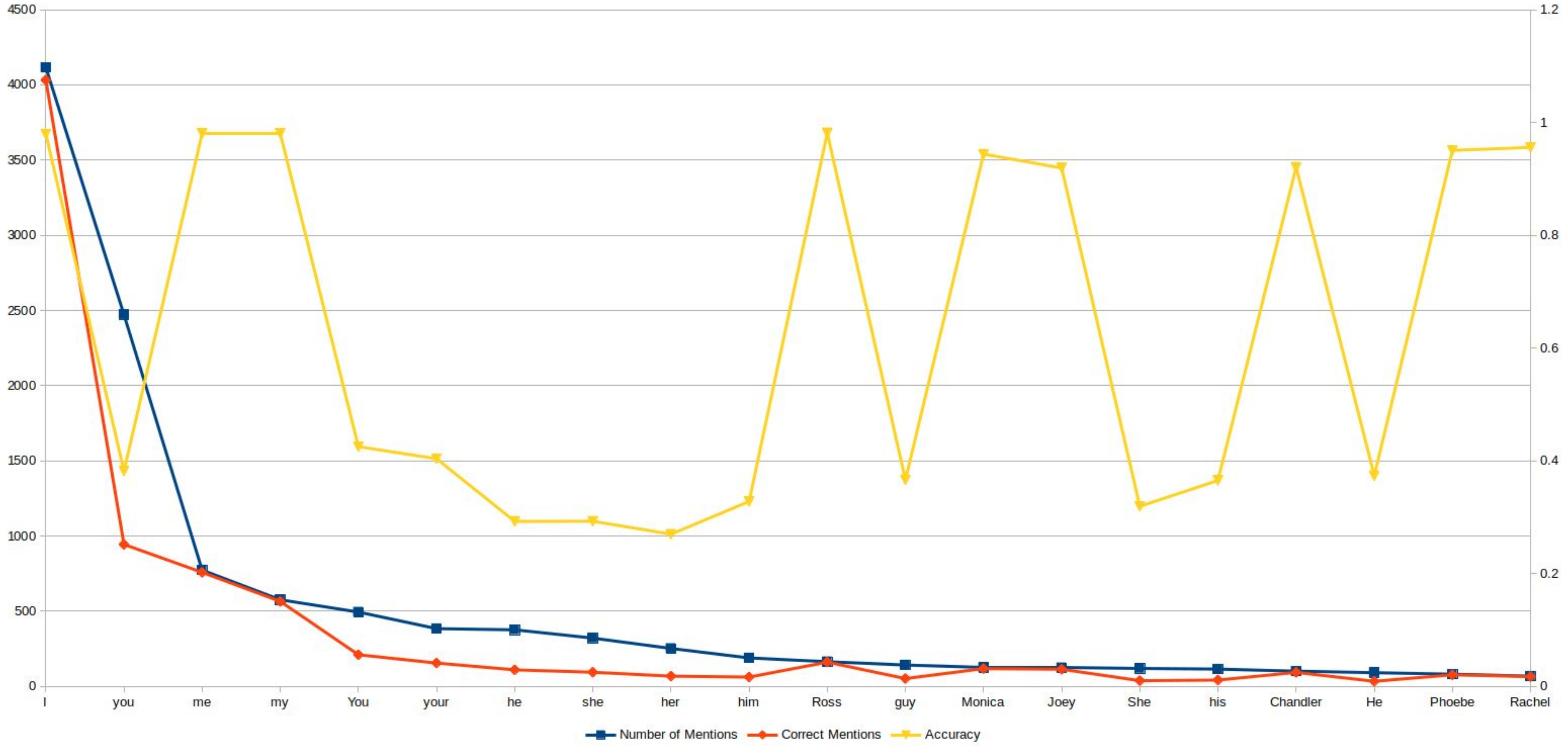
These issues are caused by data errors ignored during training. They account for about 1% of the loss.

```
correct:  (335)          our answer:(335)
correct:  (335)          our answer:(335)
correct:  (335)          our answer:(335)
correct:  (335)          our answer:(335)
correct:  (335)          our answer:(335)
correct:  (335)          our answer:(335)
correct:  (306)          our answer:(306)
correct:  (335)          our answer:(386)
correct:  (335)          our answer:(335)
correct:  (335)          our answer:(335)
correct:  (306)          our answer:(306)
correct:  (29)           our answer:(29)
correct:  (306)          our answer:(51)
correct:  (335)          our answer:(335)
correct:  (29)           our answer:(29)
correct:  (306)          our answer:(51)
correct:  (29)           our answer:(30)
correct:  (335)          our answer:(335)
correct:  (335           our answer:(215)
correct:  335)           our answer:(335)
correct:  (306)          our answer:(306)
correct:  (306)          our answer:(306)
correct:  (29)           our answer:(29)
correct:  (306)          our answer:(306)
correct:  (306)          our answer:(306)
correct:  (306)          our answer:(306)
correct:  (306)          our answer:(306)
correct:  (306)          our answer:(306)
correct:  (29)           our answer:(248)
correct:  (306)          our answer:(306)
correct:  (306)          our answer:(306)
correct:  (215)          our answer:(215)
correct:  (215)          our answer:(215)
correct:  (215)          our answer:(215)
correct:  (215)          our answer:(197)
correct:  (215)          our answer:(215)
correct:  (335)          our answer:(335)
correct:  (215)          our answer:(197)
correct:  (215)          our answer:(215)
correct:  (372)          our answer:(372)
total: 13514.0, total correct: 9772.0, accuracy: 0.723102

Process finished with exit code 0
```

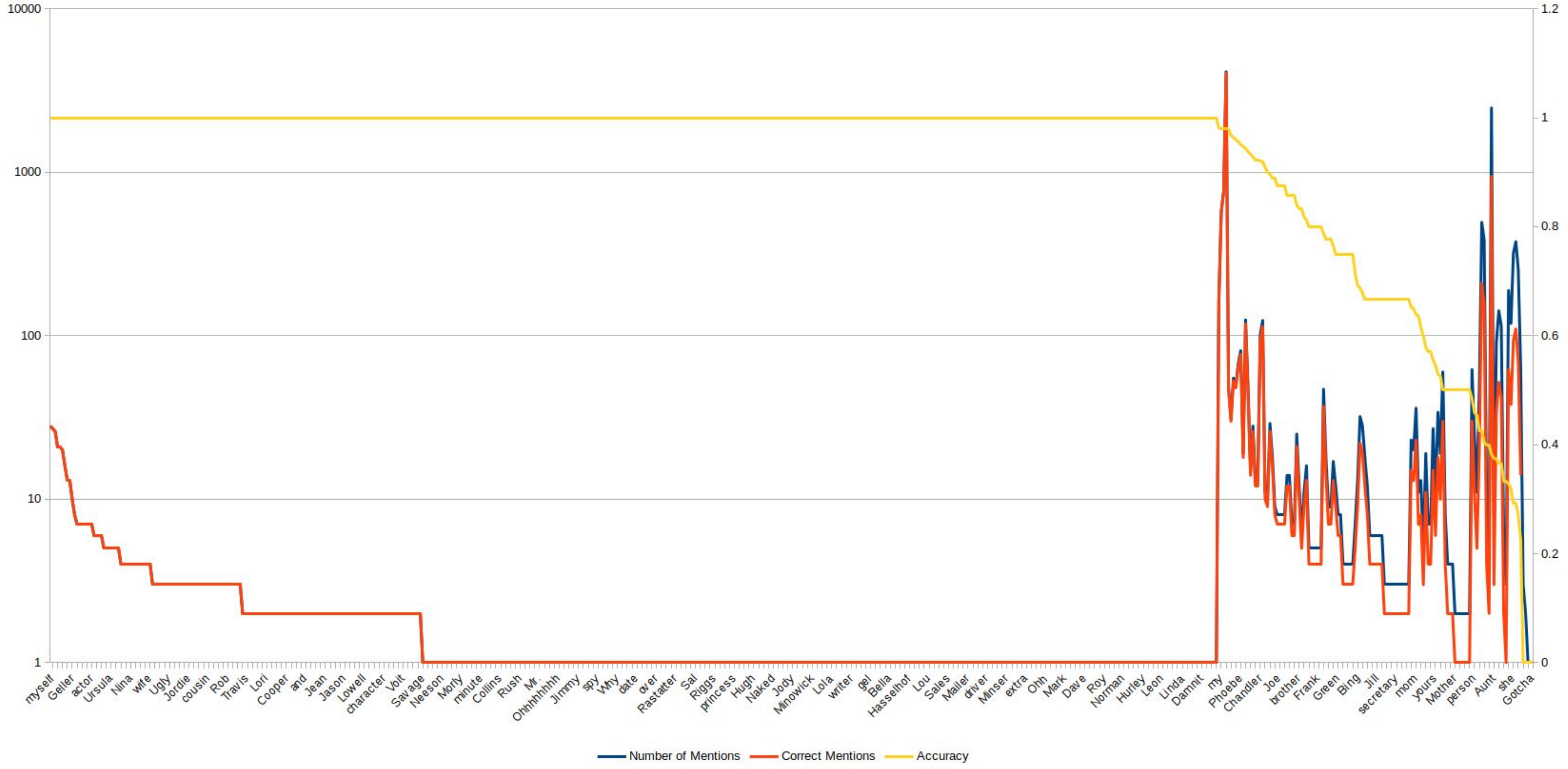

Most Mentions
# of mentions & accuracy
Accuracy
2500
2000
1500
1000
500
0
1.2
1
0.8
0.6
0.4
0.2
0
Ross_Geller
Rachel_Green
Monica_Geller
Chandler_Bing
Joey_Tribbiani
Phoebe_Buffay
Richard
Eddie
Carol_Willick
Susan_Bunch
Speaker
Number of Mentions
Correct Mentions
Accuracy



10000
1000
100
10
1
1.2
1
0.8
0.6
0.4
0.2
0
Ethan
Duncan
Frank_Buffay
Ryan
Melanie
Terry
Max
Ursula
Lipson
Janitor
Dr._Remore
Dr._Mitchell
Russ
Aunt_Iris
Best_Man
Paolo
Dr._Baldhara
Caroline
Director
Mr._Wineburg
Coma_Guy
Mr._Green
Gail
Leslie
Kid
Lorraine
Aunt_Lillian
Helen
Bob
Jeannie
Hombre_Man
Luisa
Minister
Robbie
Mrs._Wallace
Waitress
Gunther
Grandmother
Janice
Carol_Willick
Eddie
Susan_Bunch
Roger
Tilly
Erica_Ford
All
Dr._Rosen
Woman
Steve
Bernice
Mr._Tribbiani
Mr._Treeger
Nurse
Guy
Phoebe_Rachel
Teacher
Mrs._Wineburg
Waiter
Chandler_Bing
Joey_Tribbiani
Tattoo_Artist
David_Hasselhof
Flight_Attendant
Phoebe's_Assistant
Angela_Delvecchio
Joey_Phoebe_Ross
Director's_Assistant
Number of Mentions
Correct Mentions
Accuracy



4500
4000
3500
3000
2500
2000
1500
1000
500
0
1.2
1
0.8
0.6
0.4
0.2
0
I
you
me
my
You
your
he
she
her
him
Ross
guy
Monica
Joey
She
his
Chandler
He
Phoebe
Rachel
Number of Mentions
Correct Mentions
Accuracy



10000
1000
100
10
1
1.2
1
0.8
0.6
0.4
0.2
0
myself Geller actor Ursula Nina wife Ugly Jordie cousin Rob Travis Lori Cooper and Jean Jason Lowell character Voit Savage Neeson Morty minute Collins Rush Mr. Ohhhhhhhh Jimmy spy Why date over Rastatter Sal Riggs princess Hugh Naked Jody Minowick Lola writer gel Bella Hasselhof Lou Sales Mailer driver Minser extra Ohh Mark Dave Roy Norman Hurley Leon Linda Damnit my Phoebe Chandler Joe brother Frank Green Bing Jill secretary mom yours Mother person Aunt she Gotcha
Number of Mentions
Correct Mentions
Accuracy

Machine Learning Results

Naïve Bayes

Naïve Bayes performed the absolute worst of all of the machine learning algorithms we tried, though this is typical, another factor could be that all of the word2vec attributes are **not independent** of each other, meaning the independence assumption on which Naïve Bayes relies doesn't hold

```
=== Summary ===

Correctly Classified Instances         157               1.208  %
Incorrectly Classified Instances     12840              98.792  %
Kappa statistic                          0.0117
Mean absolute error                      0.006
Root mean squared error                  0.0649
Relative absolute error                107.1933 %
Root relative squared error            123.2711 %
Total Number of Instances            12997
```

SVM

SVM though cited in the literature as being one of the most promising algorithms for this task still performed rather poorly, this may be due to the class not being linearly separable with the given features and the kernel function not being sufficient to separate them

```
=== Summary ===

Correctly Classified Instances          8060               62.0143 %
Incorrectly Classified Instances        4937               37.9857 %
Kappa statistic                            0.5847
Mean absolute error                        0.006
Root mean squared error                    0.0547
Relative absolute error                  107.6734 %
Root relative squared error              103.7594 %
Total Number of Instances              12997
```

C.45 (Decision Tree)

C.45 performed the best of all the algorithms we tried, however given the tendency of decision trees to overfit and the fact that we both trained and tested on the same data this is to be expected

```
=== Summary ===

Correctly Classified Instances       10985               84.5195 %
Incorrectly Classified Instances      2012               15.4805 %
Kappa statistic                          0.8317
Mean absolute error                      0.0013
Root mean squared error                  0.0252
Relative absolute error                 22.7873 %
Root relative squared error             47.7837 %
Total Number of Instances            12997
```

C.45 (Decision Tree) 10-Fold Cross-Validation

To avoid overfitting, and to estimate how well we can expect to perform on testing data we took our best algorithm (C.45) and performed 10-Fold Cross-Validation

```
=== Summary ===

Correctly Classified Instances        9321              71.7165 %
Incorrectly Classified Instances      3676              28.2835 %
Kappa statistic                          0.6922
Mean absolute error                      0.002
Root mean squared error                  0.0363
Relative absolute error                 35.0622 %
Root relative squared error             68.8707 %
Total Number of Instances            12997
```

Which Attributes are the Most Valuable?

We use WEKA to calculate the Mutual Information for each attribute. Unsurprisingly, the *Speaker_ID* has the highest Mutual Information (Gain Ratio). This may also be an indicator for why the most likely tag baseline performed so well.

```
Ranked attributes:
 0.4298      3 speaker_id
 0.3964     62 a58
 0.3856      9 a5
 0.3787      8 a4
 0.3257     12 a8
 0.3039     50 a46
 0.2997     85 a81
 0.2729      2 episode_id
 0.2673     58 a54
 0.25        1 season_id
 ...
```

Next Steps

What we Learned from Stage 1

- The data we were given is not ideal, but we will have to manage with what we have
- Given the most likely tag baseline performs so well, any gains from machine learning will most likely be minimal
- Decision trees look very promising
- We may need more features from different sources to allow the feature vectors to become more distinct
- Until we are given proper testing data 10fold cross-validation is a must

Stage 2

- We believe that there are many more features that we are either incorrectly assessing, or missing altogether. This may be grounds for creating a neural network to more correctly classify vectors. We will explore this possibility by using Tensorflow to create a simple neural network and investigate our results.
- Since we don't want to abandon decision trees, and they performed better than our baseline, we will use them as our new baseline for our trials moving forward and explore ways that we can better improve them either by tuning parameters or adding additional features

QUESTIONS?