OUTLINE

• Example text to extract information.

• Techniques which can be used to extract that information.
  • Libraries

• How to measure accuracy.
Mr. Jack Ashley (Stoke-on-Trent, South): The hon. Member for Luton, South (Mr. Bright) made a fine speech. I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey). The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit. He was trying to have it both ways and, of course, he failed.

The hon. Member for Rugby and Kenilworth offered the House a completely bogus argument in saying that Britain has always had corporal punishment and that it is a tradition. Imagine the House debating slavery and advocates of that practice saying, "We have always had slavery in this country and it is a tradition." Imagine our predecessors in this place saying, "We have always denied women the vote, so why should we now allow women to vote? It is a tradition that they have not had the vote." I know that even now some hon. Members do not like the concept of women having the vote. Some Members prefer slavery and others prefer corporal punishment. Let it be understood that the tradition argument is bogus and nonsensical.

The hon. Member for Rugby and Kenilworth was offering us a message of despair, in stark contrast to the message offered by Mr. Martin Rosenbaum of the Society of Teachers Opposed to Physical Punishment. He has presented a marvellous message of enlightenment, in which he sets out how Britain can rid itself of corporal punishment to the advantage of teachers and pupils. I pay tribute to Mr. Rosenbaum for the fine work that he has done on behalf of STOPP. Those who oppose STOPP surely do not know anything about the research that it has conducted.

Hansard Corporal Punishment (22/07/1986):
WHAT WE WANT TO EXTRACT

• Extract relations between people automatically.
• Every mention of people as individuals
Or
• Organisations which we consider to have the same properties as a person.
WHY ONLY INDIVIDUALS?

- “Some Members prefer slavery and others prefer corporal punishment”
- We know members are being attacked by this statement by the comparison of slavery to the topic of debate corporal punishment.
- Could we definitely say which members that is though?
- Could we do that automatically?
Mr. Jack Ashley (Stoke-on-Trent, South): The hon. Member for Luton, South (Mr. Bright) made a fine speech. I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey). The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit. He was trying to have it both ways and, of course, he failed.

The hon. Member for Rugby and Kenilworth offered the House a completely bogus argument in saying that Britain has always had corporal punishment and that it is a tradition. Imagine the House debating slavery and advocates of that practice saying, "We have always had slavery in this country and it is a tradition." Imagine our predecessors in this place saying, "We have always denied women the vote, so why should we now allow women to vote? It is a tradition that they have not had the vote." I know that even now some hon. Members do not like the concept of women having the vote. Some Members prefer slavery and others prefer corporal punishment. Let it be understood that the tradition argument is bogus and nonsensical.

The hon. Member for Rugby and Kenilworth was offering us a message of despair, in stark contrast to the message offered by Mr. Martin Rosenbaum of the Society of Teachers Opposed to Physical Punishment. He has presented a marvellous message of enlightenment, in which he sets out how Britain can rid itself of corporal punishment to the advantage of teachers and pupils. I pay tribute to Mr. Rosenbaum for the fine work that he has done on behalf of STOPP. Those who oppose STOPP surely do not know anything about the research that it has conducted.
EXTRACTION TECHNIQUES

- The hon. Member for Luton, South (Mr. Bright) made a fine speech.
- I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey).
- The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit.

- Each of these sentences can be extracted in the same way.
- Requires research into the data as a whole.
The hon. Member for Luton, South (Mr. Bright) made a fine speech.

I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey).

The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit.

All extracted using rules which are specific to the topic domain.

These can be applied to any topic as long as there are common properties of the sentence.

Created by you.
• He was trying to have it both ways and, of course, he failed.
He was trying to have it both ways and, of course, he failed.

- We break a sentence down into nouns, verbs, adjectives etc.
- Search for POS we need so we can determine if we should extract the sentence.
- We can do this using a library.
- Stanford Part-Of-Speech Tagger:
  http://nlp.stanford.edu/software/tagger.shtml
STANFORD PART-OF-SPEECH TAGGER

• Sentences are broken down into individual tokens (words) and then each assigned a POS.

• Uses Penn Treebank Tag set:
  
<table>
<thead>
<tr>
<th></th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>Noun</td>
<td>Noun</td>
</tr>
<tr>
<td>NNS</td>
<td>Noun, Plural</td>
<td>Noun, Plural</td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, singular</td>
<td>Proper noun, singular</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper Noun plural</td>
<td>Proper Noun plural</td>
</tr>
<tr>
<td>PRP</td>
<td>Personal Pronoun</td>
<td>Personal Pronoun</td>
</tr>
<tr>
<td>PRP$</td>
<td>Possessive Pronoun</td>
<td>Possessive Pronoun</td>
</tr>
</tbody>
</table>

• We can look for tokens in a sentence which have been tagged as Pronouns (He, She, him, her).

• Then extract the sentence if it contains a pronoun.
• The hon. Member for Luton, South (Mr. Bright) made a fine speech.

• I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey).

• The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit.

• He was trying to have it both ways and, of course, he failed.

• The hon. Member for Rugby and Kenilworth offered the House a completely bogus argument in saying that Britain has always had corporal punishment and that it is a tradition.

• He has presented a marvellous message of enlightenment, in which he sets out how Britain can rid itself of corporal punishment to the advantage of teachers and pupils.
The hon. Member for Rugby and Kenilworth was offering us a message of despair, in stark contrast to the message offered by Mr. Martin Rosenbaum of the Society of Teachers Opposed to Physical Punishment.

I pay tribute to Mr. Rosenbaum for the fine work that he has done on behalf of STOPP.

Both sentences can be extracted by domain specific rules and POS tagging.

BUT,

Both can be extracted by something else.
The hon. Member for Rugby and Kenilworth was offering us a message of despair, in stark contrast to the message offered by Mr. Martin Rosenbaum of the Society of Teachers Opposed to Physical Punishment.

I pay tribute to Mr. Rosenbaum for the fine work that he has done on behalf of STOPP.

Each sentence can be extracted using (NER).
We can look for people in sentences and then decide if the sentence should be extracted.
Use Stanford NER:
http://nlp.stanford.edu/software/CRF-NER.shtml
STANFORD NER

• Have the use of 3 different models:

• 3 class, 4 class and 7 class.
  
  • 3 class: Location, Person, Organisations
  
  • 4 class: Location, Person, Organisation, Misc
  
  • 7 class: Location, Person, Organisation, Money, Percent, Date, Time.

• Each are trained using slightly different data so the accuracy of each model to decide what is and isn’t a person will vary.
7 class looks like the best option just because it does more.

Accuracy is a problem however and because it does more it may not have the greatest fine grained accuracy.

Use 4 class.
UPDATE

• We’ve managed to extract all the sentences we want from the text.

• Our main goal though was to extract relations between speakers.

• So we still have to:
  • Decide which sentence was said by who.
  • Who the target for each sentence was.
  • Whether a sentence is positive or negative.
WHO SAID WHAT AND WHO WAS THE TARGET

• This is solved using Anaphora resolution. Which basically means referring back to something said earlier.

• To find who said what.

• This is easy in Hansard because every statement is marked by a speaker.

• In this case it is Mr. Jack Ashley.

• Deciding the target for each sentence is much harder.
TARGET FOR SENTENCES

• The hon. Member for Luton, South (Mr. Bright) made a fine speech.
• I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey).
• The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit.

• We can use the previous techniques we devised to extract sentences and instead find the target.
• We search for hon. Member and if a location or name in a bracket is used then we have the target for our sentence.
• How do we decide if there is a location?
  • NER it extracts locations OR the use of the word “for” which is a bit easier.
The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit.

He was trying to have it both ways and, of course, he failed.

We can look back the way, if the speaker uses “He” or “hon. Gentleman” we know at an earlier point the target is referred to.

We can look back a sentence at a time and use the same set of rules for targets and then we know who the target for these sentences is.

I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey).

This precedes both sentences so the target is Mr. Pawsey.
TARGET FOR SENTENCES

• We can also extend NER and use it for Anaphora resolution.
• If a sentence contains a person’s name then the target is that person.
The hon. Member for Luton, South (Mr. Bright) made a fine speech.

**Target:** Mr. Bright  
**Source:** Mr. Ashley

I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey).

**Target:** Mr. Pawsey  
**Source:** Mr. Ashley

The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit.

**Target:** Mr. Pawsey  
**Source:** Mr. Ashley
To get ethos and decide whether or not it is positive or negative we need to perform Sentiment Analysis.

Sentiment analysis involves using features from the sentence which can then be used to define if it is positive or negative.

Two approaches:
- Discourse
- OR
- Machine Learning
DISCOURSE SENTIMENT ANALYSIS

• Keywords are used to decide if a sentence is positive or negative.

• This is done using word dictionaries with words classified as either positive or negative. – Manual.

• Bad – Negative.

• Good – Positive.

• Each sentence is tallied for its number of negative and positive words. With the most of either producing the classification to positive or negative.
Bing Lui created a dictionary of around 6,000 words:
https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

A list of positive and negative words.

Words are not plural
HOW TO AID DISCOURSE SENTIMENT ANALYSIS

• Removal of stop words. E.g. “The”, “A” etc.
  • This has been argued for and against with varying and results.
  • Depends on the data set.

• Stemming and Lemmatization
  • Involves producing the base form of a word from a plural.
  • Cats – cat
  • Cars – car
  • Use Stanford Stemming and Lemmatization:
  • Could mean that the sentiment analysis is more accurate.
PROBLEMS WITH DISCOURSE SENTIMENT ANALYSIS

• Words preceded by not. E.g. Not good, Not bad.
  • Completely changes the polarity from negative to positive and vice versa.

• Words which have not been considered E.g. Bogus.
  • An uncommon word which may not be contained in a sentiment dictionary.

• Time consuming.
  • To get accurate sentiment analysis you would need to create an exhaustive list of all good or bad words.
  • If your text is over 40,000 words it can take a long time.
MACHINE LEARNING SENTIMENT ANALYSIS

• Machine learning can be used for this purpose.

• This involves training a classifier on some manually annotated data and then running it over text to produce classifications.

• We can take a percentage of our transcript say 70% and use that as training data and then use the remaining 30% as test data.
MACHINE LEARNING SENTIMENT ANALYSIS

• Naïve Bayes and Support Vector Machines (SVM).

• Tonnes of online tutorials and tools you can use.

• SVM’s tend to produce a higher accuracy than Naïve Bayes when performing sentiment analysis. But it’s dependent on the dataset.

ML BASIC APPROACH

• To use ML we need to extract features from the text that can highlight important information.

• Most basic approach is Bag of Words
  • We take all the words for a sentence and create a dictionary from them.
  • The cat sat on the mat, is transferred into:
  • 1:2 2:1 3:1 4:1 5:1
  • When performing sentiment analysis we can decide if this sentence is positive or negative and let our SVM classifier do the work.

• This is just words but anything can be added as a feature to: length, POS etc.
ML BASIC APPROACH

• We can then extract unigrams, bigrams and trigrams and use these as features in our dictionary too.

• So we get:

• The, cat, sat, on, mat, the cat, cat sat, sat on, on the, the mat, the cat sat, etc.
PROBLEM’S USING ML

• We need to have a very large data set.
• If we have 30 sentences which we use to classify a full transcript containing 100 sentences, accuracy of the sentiment classifier will be poor.
• Again ML training is time consuming as it involves manually annotating a large amount of data.
PROBLEM OF USING ML ON OUR DATA

• The hon. Member for Luton, South (Mr. Bright) made a fine speech.
  **Target:** Mr. Bright  **Source:** Mr. Ashley

• I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey).
  **Target:** Mr. Pawsey  **Source:** Mr. Ashley

• The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit.
  **Target:** Mr. Pawsey  **Source:** Mr. Ashley
The hon. Member for Luton, South (Mr. Bright) made a fine speech.

**Target:** Mr. Bright  
**Source:** Mr. Ashley

I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey).

**Target:** Mr. Pawsey  
**Source:** Mr. Ashley

The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit.

**Target:** Mr. Pawsey  
**Source:** Mr. Ashley
PROBLEM OF USING ML ON OUR DATA

• We need to remove names, locations, domain rules so that the SVM isn’t influenced with these when classifying.

• Again need to perform Stemming and Lemmatization so that words can be matched in the dictionary.

• Again we need to decide if the removal of stop words will help in classifying data.

• All in all ML is mainly about trial and error on your data.
• The hon. Member for Luton, South (Mr. Bright) made a fine speech.

  **Target:** Mr. Bright  
  **Source:** Mr. Ashley  
  **Sentiment:** Positive

• I have never heard such a bad speech from the hon. Member for Rugby and Kenilworth (Mr. Pawsey).

  **Target:** Mr. Pawsey  
  **Source:** Mr. Ashley  
  **Sentiment:** Negative

• The hon. Gentleman tried to observe the letter of the EEC regulations, while dodging their spirit.

  **Target:** Mr. Pawsey  
  **Source:** Mr. Ashley  
  **Sentiment:** Negative
The hon. Member for Rugby and Kenilworth was offering us a message of despair, in stark contrast to the message offered by Mr. Martin Rosenbaum of the Society of Teachers Opposed to Physical Punishment.
CHALLENGES

• We can segment the sentence:

The hon. Member for Rugby and Kenilworth was offering us a message of *despair*,

*in stark contrast* to the message offered by Mr. Martin Rosenbaum of the Society of Teachers Opposed to Physical Punishment.

• But do we then lose the context of the previous sentence?

• We could use the previous sentence as a feature for classifying the next.
EVALUATING THE CLASSIFIER

• Many different metrics to evaluate the effectiveness of a classifier
• F1-score, accuracy, Kappa…
• They all have advantages
• Again depending on your dataset
EXAMPLE

• We have 100 sentences, with 20 containing a relationship between two people.

• We classify 5 correctly as having a relationship, and classify 70 as not containing a relationship.

• So 15 are classified as negative when they are positive and 10 classified as positive when they are negative.

• Use a confusion matrix to help.
## CONFUSION MATRIX

<table>
<thead>
<tr>
<th>True Positives</th>
<th>False Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>False Positives</td>
<td>True Negatives</td>
</tr>
<tr>
<td>10</td>
<td>70</td>
</tr>
</tbody>
</table>
EVALUATION

• We can calculate accuracy: \( \frac{TP + TN}{TP + TN + FN + FP} = 0.75 \)

• But we only classified 5 relations correctly yet we have 75% accuracy.

• F1-score is better for evaluating this:
  • We need precision (P): \( \frac{TP}{TP + FP} = 0.33 \)
  • We need recall (R): \( \frac{TP}{TP + FN} = 0.25 \)
  • F1-score = \( 2 \times \frac{(P \times R)}{(P + R)} = 0.28 \)
SUMMARY

• Example text to extract information.
• Techniques which can be used to extract that information.
  • Stanford Libraries for extraction
  • ML
• Some of the challenges we face.
• Evaluation of classifier