

Identification of Predictive Sentences in Political Commentary

By Richard Senington, work conducted during an internship with the University Of Leeds over the Summer of 2008

Hey, you know there's loads of raw data on the internet

Well, you know loads of it's now written by ordinary people posting to blogs and stuff

And you know their posts are loaded with opinions and predictions?

Well those posts are a reflection of public opinion about events and the future. They also influence other people as they read and hence modify overall public opinion. Eventually the future events themselves are actually influenced by the opinions being expressed.

yeah

yeah

For god sake, yes

Well duh

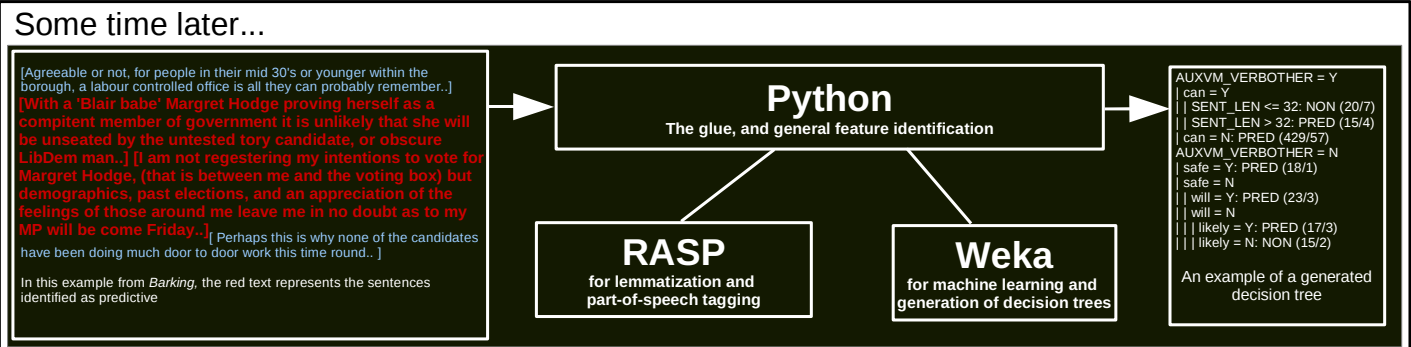
Well, wouldn't it be cool if we could automatically mine the internet for this stuff?

Hey yea, we could rig elections, rip off the stock market, win on Betfair.

OMG, I had not thought of that, so where do we start?

We need a system to identify predictive sentences, then we can try to figure out what the predictions are about afterwards.

We will need a collection of sentences to work on, we can use on-line posts discussing a major election.



Well, we've created an annotation scheme, so we can apply some rigour to our experiments. Now we just need to annotate a vast amount of text. How much do you think we should do?

Then we use the RASP lemmatizer to tag the Parts of Speech of each word in each sentence.

We'll need a way to link Weka and RASP to our other code.

We won't know for a while, we had better build the automatic system, then we can generate learning curves of how much improvement is being made, that will tell us when to stop.

That's it, then we can use our Python code to select features from the RASP output, and identify other more general features. Finally the Weka application can be run on the feature sets to identify general patterns. It will generate a decision tree to describe what it is doing and why.

No problem, Python acts as the glue, holding the entire system together.

Several months more work later...

So this learning curve says more annotation is pointless?

Yep

Well it's not bad, the error rate is 17%, from a baseline of about 50%.

Not bad at all, of course there is lots of room for improvement. RASP seems to have some problems with the noisy data we got off the internet, we need something more robust.

We could also try integrating it into a discourse system, so we could analyse transcripts of political debates, or see if there are any patterns between posts themselves.

Maybe we should try to integrate it into a practical election-forecasting system just on posts first.

I would like to thank XKCD for the stylistic inspiration to this work.