Assignment 3 - Detecting entities using the Open Calais API

Robert Walport (rpw2114) & Rashmi Raman (rr2779)

Method of creating the text corpus

We selected 5 random articles (1 each from The Economist, The Guardian, The Daily Mirror, The Independent and The Daily Mail) as our corpus. The randomness was introduced using this method :

a. Get 5 dates from an online random date generator
b. Get 5 random numbers from 1 to 20 from an online random number generator (this is to pick up the nth article from the archives for a particular day
c. Shuffle these lists using custom code randomize.py to come up with the following combinations:

1. 13th article from The Daily Mail archived on 26-Oct-2011
2. 9th article from The Independent archived on 17-Apr-2010
3. 7th article from The Mirror archived on 02-Aug-2012
4. 4th article from The Guardian archived on 18-Jun-2011
5. 10th article from The Economist archived on 10-Sep-2012

d. Copy these articles from the site (which is why we did not select sites that enforced a paywall)
e. Clean the text so that a majority of it fits in the ASCII char set (to help with string replace functions). Some text like the £ sign were left as is - these appear as '?' in the output text

Analyzing the text using OpenCalais

We wrote a small Python script to invoke the Open Calais API to analyze the text.
As per the instructions provided in the assignment:

- Skip certain entity types : we skipped IndustryTerm, Currency and MedicalCondition
- If the entity had a reference associated with it, then it was used as the link
- Else, the entity string was converted to Wikipedia format (Barack Obama became Barack_Obama) and we linked this entity to its Wikipedia page

Analyzing the results

The following 5 tables summarize our findings for the 5 texts :

The Economist

Entity	Link	Remarks
India	Detected in OpenCalais. Linked to OpenCalais	
general election	Detected in OpenCalais, but no link. Linked to Wikipedia	wrong wikipedia link – should be 2009 general election
Cable TV	Detected in OpenCalais, but no link. Linked to Wikipedia	
political journalist	Detected in OpenCalais, but no link. Linked to Wikipedia	
Sonia Gandhi	Detected in OpenCalais, but no link. Linked to Wikipedia	
Toyota	Detected in OpenCalais. Linked to OpenCalais	
London	Detected in OpenCalais. Linked to OpenCalais	
Bihar	Detected in OpenCalais. Linked to OpenCalais	
Uttar Pradesh	Detected in OpenCalais. Linked to OpenCalais	
president	Detected in OpenCalais, but no link. Linked to Wikipedia	Would've been better if it was Congress President
Aarthi Ramachandran	Detected in OpenCalais, but no link. Linked to Wikipedia	No link in Wikipedia
Rahul Gandhi	Detected in OpenCalais, but no link. Linked to Wikipedia	
leading politician	Detected in OpenCalais, but no link. Linked to Wikipedia	No link in Wikipedia
Kerala	Detected in OpenCalais. Linked to OpenCalais	
politician	Detected in OpenCalais, but no link.	

	Linked to Wikipedia	
Ramachandran's "Decoding Rahul	Detected in OpenCalais, but no link. Linked to Wikipedia	
Congress	Detected in OpenCalais, but no link. Linked to Wikipedia	Should have linked to Congress_disambiguation page instead, since Congress links to page related to the US government
Hazare	Detected in OpenCalais, but no link. Linked to Wikipedia	Takes reader to a disambiguation page
Gandhi dynasty	Not detected	
Rural affairs minister	Not detected	
Pronouns	Not detected	9 instances of He, 1 instance of she, 1 instance of her, 6 occurrences of his, 3 occurrence of their,
Entities detected	**18**	4 were not detected but not considered as the type was in our filter
Total entities	**37**	
Correct links	**13**	
% Entities detected	**48%**	
% Correct links	**72%**	

Entity	Link	Remarks
head waiter	Detected in OpenCalais, but no link. Linked to Wikipedia	Links to a page that has been deleted in Wikipedia
Norwich	Detected in OpenCalais, but no link. Linked to Wikipedia	
Alex Tranquillo	Detected in OpenCalais, but no link. Linked to Wikipedia	Page does not exist in Wikipedia
David Adlard	Detected in OpenCalais, but no link. Linked to Wikipedia	Page does not exist in Wikipedia
Bishop	Detected in OpenCalais, but no link. Linked to Wikipedia	Takes reader to the wrong page
chef	Detected in OpenCalais, but no link. Linked to Wikipedia	
United Kingdom	Detected in OpenCalais. Linked to OpenCalais	
London	Detected in OpenCalais. Linked to OpenCalais	
US Federal Reserve	Detected in OpenCalais, but no link. Linked to Wikipedia	Page does not exist in Wikipedia
Food items	Not detected	smoked ham hock terrine , spiky piccalilli,toast,poached egg,Norfolk lamb, butter fondant potatoes, jus,samphire ,sea trout,baked apple,toffee sauce,chocolate brownie
Adlard's	Not detected	
Walpole Arms	Not detected	
Itteringham	Not detected	
Ramsay Savoy Grill	Not detected	
pronouns	Not detected	5 instances of they, 1 instance of he
Entities detected	**9**	

Total entities	**27**	
Correct links	**4**	
% Entities detected	**33**	
% Correct links	**44**	

The Mirror

Entity	**Link**	**Remarks**
Paris	Detected in OpenCalais. Linked to OpenCalais	
France	Detected in OpenCalais. Linked to OpenCalais	
Didier Drogba	Detected in OpenCalais, but no link. Linked to Wikipedia	
Stamford Bridge	Detected in OpenCalais, but no link. Linked to Wikipedia	
Twitter	Detected in OpenCalais. Linked to OpenCalais	
Ashley Cole	Detected in OpenCalais, but no link. Linked to Wikipedia	
Salomon	Detected in OpenCalais. Linked to OpenCalais	Salomon has the wrong link because Salomon was considered as an entity, not Salomon Kalou
Carlo Ancelotti	Detected in OpenCalais, but no link. Linked to Wikipedia	Page does not exist in wikipedia
United Kingdom	Detected in OpenCalais. Linked to OpenCalais	
Ryan Bertrand	Detected in OpenCalais, but no link. Linked to Wikipedia	

Ancelotti	not detected	
Saint-Germain	not detected	
Chelsea	not detected	
PSG	not detected	
Manchester United	not detected	
Getty	not detected	
pronouns	not detected	7 instances of his, 1 instance of them, 1 instance of they
Entities detected	**10**	
Total entities	**25**	
Correct links	**8**	
% Entities detected	**40**	
% Correct links	**80**	

The Independent

Entity	**Link**	**Remarks**
Amy Jenkins	Detected in OpenCalais, but no link. Linked to Wikipedia	
Nick Clegg	Detected in OpenCalais, but no link. Linked to Wikipedia	
America	Detected in OpenCalais, but no link. Linked to Wikipedia	Takes reader to a disambiguation page
John Lewis	Detected in OpenCalais, but no link. Linked to Wikipedia	Takes reader to a disambiguation page

therapist	Detected in OpenCalais, but no link. Linked to Wikipedia	
Friends	Detected in OpenCalais, but no link. Linked to Wikipedia	Correct link to the TV show
Monica	Detected in OpenCalais, but no link. Linked to Wikipedia	Takes reader to a disambiguation page
Rachel	Detected in OpenCalais, but no link. Linked to Wikipedia	Takes reader to a disambiguation page
New York	Detected in OpenCalais. Linked to OpenCalais	
Jung	Not detected	
Freud	Not detected	
Clegg	Not detected	
Brown	Not detected	
Cameron	Not detected	
The Catcher in the Rye	Detected as Catcher	Wiki link to Baseball Catcher
Salinger	Not detected	
Holden Caulfield	Not detected	
Punch and Judy	Not detected	
scriptwriters	Not detected	
worm	Not detected	
leaders	Not detected	As in Cameron, Brown and Clegg
Pronouns	Not detected	3 instances of its, 4 instances of he, 7 instances of they, 10 instances of I, 2 instances of me, 6 instances of we, 3 instances of them
Entities detected	**10**	
Total entities	**59**	
Correct links	**5**	

% Entities detected	**16.9**	
% Correct links	**50**	

The Daily Mail

Entity	**Link**	**Remarks**
Marks & Spencer	Detected in OpenCalais. Linked to OpenCalais	
Marc Bolland	Detected in OpenCalais, but no link. Linked to Wikipedia	
Frazer Ramzan	Detected in OpenCalais, but no link. Linked to Wikipedia	Page does not exist in wikipedia
analyst	Detected in OpenCalais, but no link. Linked to Wikipedia	Takes reader to a disambiguation page
Bolland	Not detected	
Nomura	Not detected	
M&S	Not detected	
Marks and Spencer	Not detected	
clothing retailers	Not detected	
Analysts	Not detected	
Next	Not detected	(the clothing retailer)
Pronouns	Not detected	4 instances of its, 2 instances of he,
Entities detected	**5**	
Total entities	**18**	
Correct links	**3**	

% Entities detected	**27.7%**	
% Correct links	**60%**	

Observations

Persons

The Open Calais API is quite accurate in detecting if an entity is a person when they have a first and surname - all the results show consistently that Open Calais detects these full names quite well. Single surnames out of context (such as Freud and Salanger) it tends to fail to detect in almost every case.

Places

Open Calais is extremely accurate at detecting geographical places - and linking them to the Open Calais DB. It also can detect where disambiguation is necessary - for example Paris.

Companies

Open Calais also is pretty accurate in detecting companies - and linking them to the Open Calais DB. Having said that, we are not entirely sure if Open Calais has a higher priority for detecting companies or persons first - this might be quite a contentious issue in the case of firms which are generally named after their partners/majority shareholders like law firms.

Government

Open Calais can detect government organizations and political parties too - but they do have a U.S. context, which can lead to incorrect results.

Pronouns

Open Calais does not detect pronouns at all - this might be a useful feature to have for entity disambiguation.

Occupations

Open Calais has a pretty good accuracy in detecting occupations and positions.