Detailed Text Categorisation for Wikipedia

Author:Sam Tardif, star4245@it.usyd.edu.au
Supervisors: Dr James R Curran and Dr Tara Murphy
School of Information Technologies

1. Project Aims

We have been exploring Wikipedia as an exciting new resource for text categorisation. Wikipedia provides a unique set of rich additional content that we have utilised for the task of classifying its articles. While the applications of a high quality labelled set of Wikipedia articles are numerous, an improvement in Named Entity Recognition (NER), the task of classifying named entities (proper nouns) in text, was our motivating goal.

2. Wikipedia as a Corpus

- Wikipedia is free, huge (~2,800,000 articles), and packed with information, like infoboxes and categories, that's useful for categorising articles

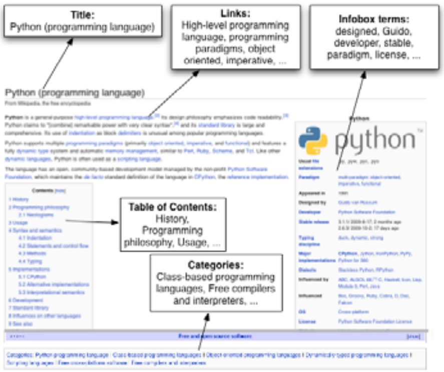

Title:
Python (programming language)
Links:
High-level programming language, programming paradigms, object oriented, imperative, ...
Infobox terms:
designed, Guido, developer, stable, paradigm, license, ...
Python (programming language)
python
Table of Contents:
History, Programming philosophy, Usage, ...
Categories:
Class-based programming languages, Free compilers and interpreters, ...

Figure 1. A sample Wikipedia article and some of its Interesting components

3. Categorising Wikipedia

- Using *mwlib* we extracted from Wikipedia's rich article content a set of features to represent each document
- In line with our applied task of NER, we classified articles as being about a person, organisation, location, miscellaneous entity or common noun (and further subcategories) using the Naïve Bayes (NB) and Support Vector Machine (SVM) algorithms

4. But First Some Annotation

- Accurate, manually annotated data is required to train a machine learner
- We developed a tool that tracks annotation statistics, caters for the preferences of the annotator, and can be used to tune the annotation process

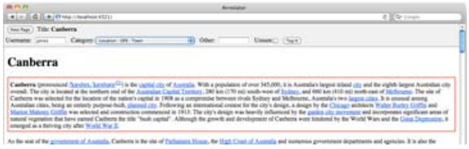


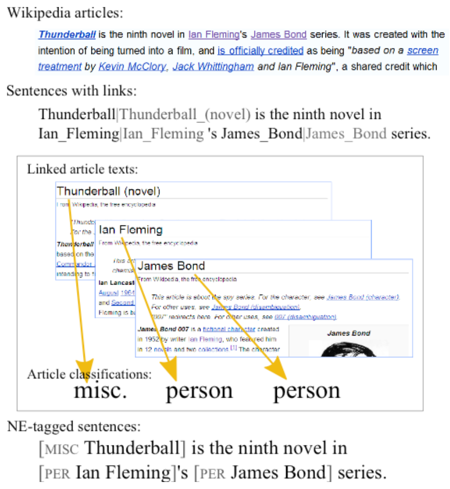
Figure 2. The annotation tool displaying a Wikipedia article

5. Who or what is Paris Hilton?

- Consider the following two sentences:

> *I'm going to visit Paris Hilton today.*
>
> *I'm going to visit **the** Paris Hilton today.*

- This ambiguity is one of the core problems of NER
- Assuming one has identified both instances of "Paris Hilton" as named entities, they would then need to determine which refers to a person, location or organisation
- Wikipedia's link structure to the rescue!



Wikipedia articles:

Thunderball is the ninth novel in Ian Fleming's James Bond series. It was created with the intention of being turned into a film, and is officially credited as being "*based on a screen treatment by Kevin McClory, Jack Whittingham and Ian Fleming*", a shared credit which

Sentences with links:

Thunderball|Thunderball_(novel) is the ninth novel in Ian_Fleming|Ian_Fleming 's James_Bond|James_Bond series.

Linked article texts:

Article classifications:

NE-tagged sentences:

[MISC Thunderball] is the ninth novel in [PER Ian Fleming]'s [PER James Bond] series.

Figure 3. The complete process outlined by Nothman et al. (2009)

- Nothman et al. (2009) extracted from Wikipedia articles sentences that contain links to other Wikipedia articles
- Assuming that most articles describe objects, these links were taken as potential named entities and classified with the label given to the article to which they point

6. The Importance of Being Accurate

- Clearly the NER system relies on accurate classifications for Wikipedia articles - errors at this stage are unrecoverable!
- For an NER system to determine who or what Paris Hilton is, our text categorisation system must accurately classify the articles on **Paris Hilton** (the person), **Paris** and **The Hilton** as being about a person, location, and organisation, respectively

7. Results

- We evaluated our system on 2,311 hand-labelled documents using 10-fold cross-validation. Results for both the NB and SVM classifiers are presented. We compare our results with previous approaches to text categorisation by Nothman et al. (2009) and Dakka and Cucerzan (2008).

	NB			SVM		
Class	P	R	F	P	R	F
PER	69	99	82	99	92	96
ORG	72	94	81	95	91	93
LOC	97	99	98	99	99	99
MISC	71	83	76	90	88	89
NON	99	58	73	91	96	93
DAB	87	99	92	98	99	98
	Micro F-score: **84**			Micro F-score: **95**		

Table 1. Precision, Recall and F-score for the individual classes, as well as the micro F-score over all classes, using best NB and SVM configurations

	NB			SVM		
Class	P	R	F	P	R	F
PER	88	98	93	98	94	96
ORG	88	93	91	97	94	96
LOC	99	99	99	99	99	99
MISC	95	84	89	92	97	94
	Micro F-score: **94**			Micro F-score: **97**		

Table 2. Results for Table 1 experiments using NE categories only

Nothman	Dakka	Baseline	Full
91	90	94	**95**

Table 3. Comparison of micro F-scores

- In summary, using an SVM classifier and our feature set we were able to outperform previous state-of-the-art results for classifying Wikipedia articles

8. Future Work

- We are currently in the process of evaluating the complete NER task using our text categorisation system
- We would like to expand our annotation set to better represent all article categories
- Analysis of categorisation schemes and what is suitable for Wikipedia data

References

W Dakka and S Cucerzan. Augmenting wikipedia with named entity tags. In Proceedings of IJC- NLP 2008, 2008.

Joel Nothman, Tara Murphy, and James R. Curran. Analysing Wikipedia and gold-standard corpora for NER training. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 612-620, Athens, Greece, March 2009. Association for Computational Linguistics.

PediaPress. mwlib MediaWiki parsing library. http://code.pediapress.com.

The University of Sydney