Project Text Analysis – Final Project: Wikification

Name: Leonardo Losno Velozo
Student: 1668501
11-06-2015

The system that we developed is concerned with *wikification*. This means that it detects entities of interest in a given text and links them to a Wikipedia page. The entities of interest are: country/state (COU), city/town (CIT), natural places (NAT), person (PER), organization (ORG), animal (ANI), sport(SPO) and entertainment (ENT). The text we are using, is a 50% portion of all the hand annotated articles from North-American newspapers. Finally, the target database that is used is Wikipedia. Summarized, this system will take care of detecting the entities of interests of newspaper articles and link them to relevant Wikipedia pages.

The basic idea of this system is based on machine learning. For this we use the Stanford Named Entity Recognizer (NER) as base. This implementation finds and labels sequences of words in a text which are names of things. The default classifier is trained particularly for 3 classes (PERSON, ORGANIZATION, LOCATION). We therefore decided to train the classifier on the entities of interest as described above. The first step is to train the system with the annotated data so that it can recognize entities from our categories. Therefore we split the data in a 80% train file and a 20% test file. The train data is a separated file in which each line includes a token and a corresponding entity tag, remarking that unannotated data is given Other (O) as entity tag. When the system is trained with this data, the second step can be taken. In this step the test data is processed line by line. In this process the words of this annotated text are extracted, to remain with the raw text of these newspaper articles. These words are then tagged based on the trained classifier, described in step one. The third step is to link each tagged word to a Wikipedia page. For this step we use a Wikipedia API which takes the words that are a named entity and puts it in an URL which automatically searches for appropriate Wikipedia URL's. Therefore, following tagged words with the same tag (excluding the O tag) are taken together as a named entity separated by a an underscore, for example: 'Buenos CIT' 'Aires CIT' = Buenos%20Aires. This named entity (e.g. Uruguay) is added to the URL as a query:

'http://en.wikipedia.org/w/api.phpaction=query&list=search&srsearch=**Uruguay**&format=json'

The API gives the possible Wikipedia URL's back, where we take the first URL as a match, because this is the most wanted page. In the last step the resulting system tags and links are added to a raw copy of the articles (without the hand annotation). Also a HTML page is generated, in which each document gets its own page and each tagged entity is made clickable link to the corresponding Wikipedia page. Finally the system results the test data are checked against the results of the hand annotated test data. In here we can see that the results of the total system are: accuracy 51%, precision 63%, recall 84% and f-score 72%. For the total summary of the measure results, see the reference on the next page. The results make clear that the system can be improved, but with this we made a good first step for a *wikification* system in which most of the words will be correctly linked.

For the most part we worked together to produce this final project: we thought together finding solutions for the problem and implemented it together in a program. In the last Chris took a leading role because of his high level of programming. Therefore I learned a lot of new programming skills, beside of what I learned from the course Project Text Analysis.

Reference

Table 1: Confusion matrix

	CIT	COU	NAT	O	ORG	PER
CIT	<5>	2	.	1	.	2
COU	.	<17>	.	7	.	.
NAT	.	.	<.>	4	.	.
O	.	1	.	<507>	6	.
ORG	.	.	.	3	<1>	.
PER	1	1	.	7	.	<9>

Table 2: Precision, recall and f-score

	Precision	Recall	f-score
COU	0.81	0.71	0.76
CIT	0.83	0.50	0.62
NAT	0	0	0
PER	0.82	0.50	0.62
ORG	0.14	0.25	0.18
ANI	0	0	0
SPO	0	0	0
ENT	0	0	0
TOTAL	**0.63**	**0.84**	**0.72**