Authorship Attribution in Greek Tweets Using Author's Multilevel N-gram Profiles

George K. Mikros & Kostas Perifanos

[1]Department of Italian Language and Literature, [2]Departement of Linguistics
National and Kapodistrian University of Athens - Greece

National and Kapodistrian
University of Athens - Greece

ABSTRACT

The aim of this study is to explore authorship attribution methods in Greek tweets. We have developed the first Modern Greek Twitter corpus (GTC) consisted of 12,973 tweets crawled from 10 Greek popular users. We used this corpus in order to study the effectiveness of a specific document representation called Author's Multilevel N-gram Profile (AMNP) and the impact of different methods on training data construction for the task of authorship attribution. In order to address the above research questions we used GTC to create 4 different datasets which contained merged tweets in texts of different sizes (100, 75, 50 and 25 words). Results were evaluated using authorship attribution accuracy both in 10-fold cross-validation and in an external test set compiled from actual tweets. AMNP representation achieved significant better accuracies than single feature groups across all text sizes.

CONTACT

George K. Mikros
Department of Italian Language and Literature
National and Kapodistrian University of Athens
Email: gmikros@isll.uoa.gr
Website: http://users.uoa.gr/~gmikros

Aims of the research

- To perform authorship attribution experiments in tweets written in Modern Greek.
- Create the first Modern Greek Tweets Corpus (GTC) in order to use it as a reference corpus for studying social media language including authorship attribution, sentiment analysis and linguistic variation.
- Explore the effectiveness of a specific document representation called Author's Multilevel N-gram Profile (AMNP), which comprises of a combined vector of increasing size and different level n-grams
- Investigate alternative ways to construct training sets for authorship attribution in Twitter data.

The Greek Twitter Corpus -GTC

We compiled a new corpus of tweets written in Modern Greek (see Table 1). In order to extract tweets from the specific users we used the twitteR R package. During the corpus preprocessing we removed all @replies, #tags and manual retweets (RT's).

Authors	No of Tweets	Total size (words)	Average size (words)	Standard Deviation
A	500	5,378	10.75	5.42
B	918	10,515	11.45	5.52
C	2,065	32,098	15.54	6.73
D	455	7,451	16.57	5.48
E	1,347	9,822	7.29	5.01
F	535	3,692	6.90	4.93
G	1,277	9,412	7.37	5.63
H	2,306	26,212	11.36	5.86
I	2,986	18,720	6.26	4.28
J	584	7,618	13.06	6.74
Total	12,973	130,918		

Table 1: Descriptive statistics of GTC

Author's Multilevel N-gram Profile

Taking into consideration the complementary nature of character and word level information, we propose a combined vector of both character and word n-grams of different size.
We extracted the 1,000 most frequent character and word n-grams with n=2 and 3 resulting in a total vector of 4,000 features.
The resulting vector represents the Author's Multilevel N-gram Profile (AMNP), a document representation that captures in a parallel way both character and word sequences.

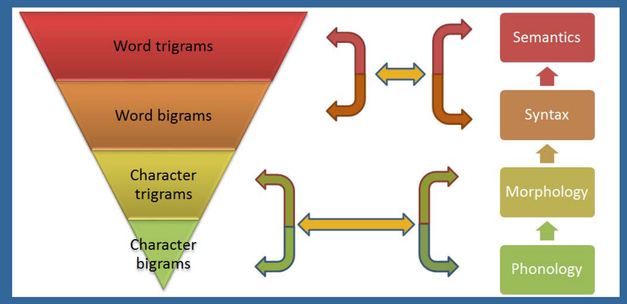

Figure 1: AMNP: An hierarchical representation of n-gram features and related linguistic levels.

Experimental Methodology

Experimental set-up:
We used GTC to create 4 different datasets which contained merged tweets in texts of sizes (100, 75, 50 and 25 words). We tested the authorship attribution accuracy with each feature group separately and compared it with AMNP. Accuracy figures were calculated on two different conditions:

a) 10-fold cross-validation (cv) in the merged tweets text units
b) External test set which contained 500 single tweets not included in the training set (35-60 per author).

Results

Authorship attribution in Greek tweets can be performed with remarkable accuracy when we use a training set in which the basic text units contain merged tweets. Best results were obtained using 100-word and 75-words text chunks (0.952 and 0.918 respectively) and using a 10-fold cross-validation procedure. However. when we used the external tweets as test set accuracy rates moved to the opposite direction with smaller text-size chunks producing better attribution rates than bigger ones (see figure 2).

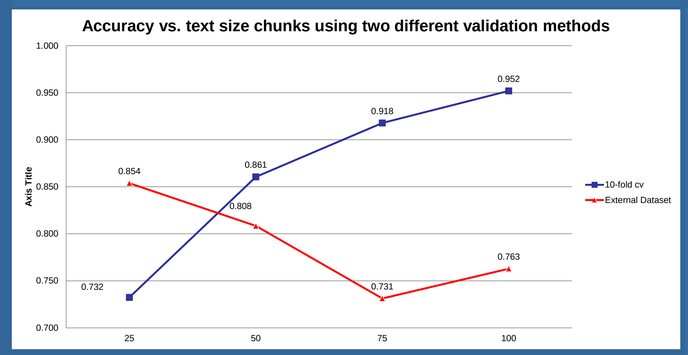

Figure 2: Impact of text size of merged tweets in the authorship attribution accuracy (10-fold cv & external dataset) .

A second experiment was conducted in order to evaluate whether AMNP representation captures better the stylometric profile of the tweets than using separate n-gram profiles. For this reason we repeated the authorship attribution task in the four datasets of varying text size chunks under both testing conditions (10-fold cv and external tweets). The results appear to the following figures (figure 3 & 4).

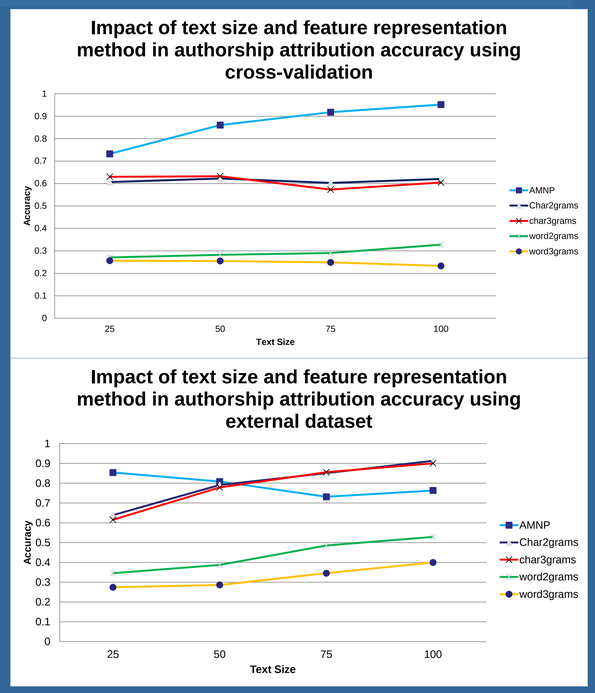

Figure 3 & 4: Impact of text size of merged tweets in the authorship attribution accuracy (10-fold cv & external dataset) .

CONCLUSIONS

- Authorship attribution in tweets of Modern Greek is a feasible task. Our top performance (0.951 accuracy in 10-fold cv using 100-word text chunks) is a good indication that the tweet's linguistic structure is a significant carrier of authorship information.
- AMNP representation proved highly efficient compared to single n-gram feature groups in all text sizes.
- Optimal results are achieved when both training and testing sets for authorship attribution contained merged tweets.
- The text-size threshold for using AMNP seems to be the 50 words per text chunk.