

Automatic Authorship Detection

Gökhan Tür

Department of Computer Engineering

Bilkent University, Bilkent, Ankara, TR-06533, TURKEY

`tur@cs.bilkent.edu.tr`

Abstract

This paper presents a statistical approach for detecting the authors of newspaper editorials. The method is based on unigram language models developed for each of the selected 9 authors of the Turkish Milliyet newspaper. When an unseen editorial is given, the system chooses the writer, whose corresponding language model maximizes the probability of that piece of text. All the training and test material were downloaded from the web archives of the newspaper. The results indicate that it is possible to reach an accuracy of above 90% using this simple method. We have also seen that stemming indeed decreases accuracy since it results in losing very important stylistic information.

1 Introduction

Automatic authorship detection systems aim to assign authors to texts. The historical roots of this work lie on anonymous or ambiguous texts, such as detecting whether John Smith or the God has been the real writer of the Mormon scripture, without any computer aid. The details of the previous work can be found in the next section.

In order to evaluate our system, we have preferred to detect the authors of texts, whose authors are already known. For this purpose, we have used real-world text, newspaper editorials, downloaded from the web, and selected 9 authors, whose detailed information is given in Table 3.

We have not employed any stylistic features, such as using their average word length, sentence length, word spectrum, etc., but instead only used the word frequencies. The details of our system is explained in Section 3. Section 4 presents analytical experimental results.

2 Previous Work

It would be useful to analyze the previous work in two phases. The early phase, which is dominated by the linguists and stylometrists, has been focused on literary detective work. In general their aim has been to distinguish two or three authors, imitating each other, or try to prove that a portion of an authors' writing does not belong to him/her. There has been no quantitative evaluation, instead they have tried to determine the criteria, which would be useful in stylometry. The details of this early work can be found in Oakes [1998].

The work of Yule has focused on developing a metric to differentiate authors. He produced and used a measure of "vocabulary richness", called the *K characteristic*. Basically, this metric indicates the probability that any randomly selected pair of words will be identical.

Later, in 1962, Ellegard has proposed to use a metric called *distinctiveness ratio*. This metric indicates how far you are from the average usage of a word. For example if the probability of the word *minister* is 0.0005, and a specific author uses it with probability 0.0001, then this ratio is $0.0001/0.0005 = 0.2$. He selected about 50 such important words, and made his analyses according to them without using any computer aid.

In 1978 Morton has found that the largest group of words in the vocabulary of a text is the words which occur only once, the *hapax legomena*. Such obscure, out-of-date, or specialist words (such as *binaenaleyh*) can help authorship detection systems. The problem with such words is that it is impossible to perform statistical analysis, since they are very infrequent in nature. The second cue, Morton has suggested, is the placement of words. He proposed to use first, second, or last words of a sentence, or words immediately before and after a certain stopword for inflected languages like Turkish or Greek, and use collocations for uninflected languages like English.

Detection of the writers of the *Federalist papers*, which were published in newspapers in 1787-1788 by 3 authors, has long been interested many researchers. Mosteller and Wallace has used Bayesian statistics for this task in 1963. Tweedie, Singh, and Holmes have used Neural Network approach in 1994. One year later, Holmes and Forsyth have used genetic algorithms in order to assign these letters to authors.

In 1966, Milic has proposed to use the number of part-of-speech trigrams, sentence-initial part-of-speech, and the use verbals in author attribution task. In a later study, Köster analyzed 4 other writers, and found out that these 3 criteria may be highly irregular.

In 1976, Kjetsaa has used average sentence length, sentence length spectrum, six part-of-speech categories, and part-of-speech in first two, and last three positions.

Note that, these studies all have focused on resolving an ambiguity, such as detecting

the writers of the Federalist papers, New Testament Epistles, Mormon Scriptures, Junius Letters, or authors of the novels “The Imitation of Christ”, Aristotelean “Ethics”, Sholokhov’s “Was and Quiet Flows the Don”, etc. So, it is impossible to compare these work with each other. For example, in 1995, Ellison showed that, according to Morton’s criteria, his own essays must each have been written by somebody else!

In a more sound study, in 1996 Baayen, van Halteren, and Tweedie employed syntactic information with a syntactically annotated corpus. They have found that the syntactic category (part-of-speech), the syntactic function (e.g. NP Head), and additional attributes (e.g. singular) can be used in authorship attribution task. They have trained and tested their system on 20 texts written by two authors and compared previous metrics, such as Yule’s K metric with theirs. They concluded that they have outperformed previous metrics. The most important drawback of their system is the requirement of expensive syntactic annotation.

In 1999, Stamatatos *et al.* have proposed to use style markers, such as sentence count, word count, noun phrase count, etc. They have not employed word frequency counts or other lexically-based measures. They have tried to assign editorials to one of the 10 authors. They have reached 69% accuracy.

3 Approach

Our approach is based purely on word distributions of the authors. Given enough training data for an author, we can build unigram language models. Then, when an unseen test editorial is given, we can select the author, whose model maximizes the probability of that editorial. More formally, we will choose the author A_i maximizing the probability $P(A_i|W)$ in the text:

$$\operatorname{argmax}_{A_i} P(A_i|W)$$

where W is the text waiting to be classified. Since W consists of words, w_j , and we are dealing with unigram language models, we can express this probability as follows:

$$P(A_i|W) = \prod_{w_j} P(A_i|w_j)$$

The only problem is to determine this probability. In order to do this, we modeled the word frequencies of the authors in different ways, as described in the following subsections.

The issue in such a formalism is determining the probability of an unknown word. This problem has been solved by smoothing techniques years ago, but they all consider only

Author	Words
AG	eğitim (education), yıl (year), 8, üniversite (university), milli (national)
AT	Beşiktaş, gol (goal), Fenerbahçe, Beşiktaş'ın (genitive case), takım (team)
DS	Refah (Welfare), Türkiye (Turkey), Yılmaz, CHP, seçim (election)
DA	kadın (woman), şey (thing), diyor (saying), insanlar (people), güzel (beautiful)
FB	Başbakan (prime minister), Yılmaz, Ecevit, Sayın (Mr.), Bakanı (minister+Poss)
HP	yıl (year), vardır (exists), bizim (our), şey (thing), Türk (Turk), olur (does)
MA	diyor (says), yıl (year), Türk (Turk), Başkanı (President), Genel (General)
SK	Türkiye'nin (Turkey's), AB (EU), Türk (Turk), Türkiye (Turkey), Kıbrıs (Cyprus)
TA	Türkiye (Turkey), RP, diyor (says), Yılmaz, Türkiye'nin (Turkey's)

Table 1: The most common words of the authors.

one language model. When we are comparing the probabilities obtained from different language models, we could not use such techniques, instead we left this probability as a parameter of our system, which can be optimized using the development data.

3.1 Word-based Modeling

In this method, we have built unigram language models from the surface forms of the words. This was the simplest approach, and the only thing we need was a tokenizer. Thinking that, the stopwords contain no author dependent information, we ignored them.

In order to give the flavor of the models, we present the top 10 most common words of the authors in Table 1.

Looking from the reverse side, we can get the probabilities for the text piece.

“Milli Eğitim Bakanlığı 8 yıllık temel eğitim projesini uyguluyor”
(literally, “Ministry of Education performs the 8 year basic education project”)

which is a typical sentence of Abbas Güçlü, who writes on education issues, and especially on the project of 8 year basic preliminary education. Note that the sports writer Attila Gökçe and Duygu Asena, who writes on woman issues are very far from writing such a sentence as seen from Table 2, whereas, other writers, writing on general or political issues get almost equal probabilities, except Fikret Bila, who must have written editorials related to this subject.

Author	Milli	Eğitim	Bakanlığı	8	yıllık	eğitim	projesi	uyguluyor	LogProb
AG	32.30	40.72	12.59	40.72	17.47	60.94	0.30	0.10	-27.10
AT	4.29	0.78	0.00	0.00	3.11	0.00	0.00	0.00	-40.97
DS	3.98	0.90	1.80	9.22	5.97	5.07	0.36	0.00	-33.29
DA	1.25	2.08	0.00	0.83	1.66	9.14	1.25	0.00	-36.05
FB	11.10	6.48	5.32	9.71	12.26	13.18	0.46	0.23	-29.72
HP	3.53	1.76	3.41	1.87	3.97	2.98	0.77	0.11	-32.95
MA	3.89	3.67	5.34	7.06	5.52	5.43	0.77	0.27	-31.20
SK	0.61	0.85	3.54	3.42	2.44	2.44	1.22	0.12	-33.84
TA	3.38	1.75	1.50	3.38	3.75	5.62	0.88	0.13	-32.66

Table 2: The probabilities ($\times 10^{-4}$) for each word and author. Final column indicates the logarithm of the overall probability for this sentence.

3.2 Stem-based Modeling

Morphologically rich languages, like Turkish, suffer from the fact that the number of word forms one can derive from a Turkish root form may be in the millions [Hankamer, 1989]. Because of this reason, the number of distinct word forms is much larger than that of English. Using the surface forms of the words results in data sparseness in the training data. When we consider the words with different inflectional and derivational suffixes different, then we have to deal with data sparseness. It is intuitive that, taking out the suffixes the words, and using the root words will prevent the data sparseness, and the unigram language models would be more effective. For example, Attila Gökçe has used various forms of the football team *Beşiktaş* in his writings, such as *Beşiktaş'ın*, *Beşiktaş'ta*, *Beşiktaş'ı*, or *Beşiktaş'a*. So we decided to use the root words instead of the surface forms of the words, and build stem-based language models, instead of word-based language models.

In order to do this, we used a preprocessing module, developed by Hakkani-Tür [2000 forthcoming], which tokenizes the training data, analyzes the tokens using the morphological analyzer developed by Oflazer [1993], groups the collocations, and finally removes some obviously improbable morphological parses in order to reduce the morphological ambiguity. Then, we extracted the roots of the words, and rebuilt the training corpus using only these roots. When there were more than one root for a word, we used all of the roots. However, this ambiguity was not a real problem as there were only 1.15 distinct roots per word on the average.

3.3 Noun-based Modeling

It can also be intuitive that, nouns would be more informative than any other types of words, thus all the words except nouns can be ignored as if they are stopwords. Since we have the morphological analyses of the words, it was straightforward for us to test this hypothesis. Instead of using the stems of words, we only used the stems of the morphological parses that have a noun root form.

4 Experiments and Results

To evaluate our models we carried out experiments with real world data. We first describe our training and test data, then give results obtained with the baseline word-based, stem-based, and noun-based language models. We used SRILM toolkit for building the language models [Stolcke, 1999].

4.1 Data

For training, we have used the editorials written by 9 different authors, in our corpus of about 18 million words, compiled from the web resources of Milliyet newspaper articles, covering the period from January 1, 1997 through September 12, 1998. This makes a total of about 725,000 words excluding the stopwords.

In order to tune the parameters of my system, we have used a development set of 254 editorials, including 136,000 words excluding stopwords, and evaluated this algorithm using an unseen test set of 255 editorials, including about 137,000 words, excluding the stopwords. The details of the training, development, and test data is given in Table 3.

4.2 Evaluation Metrics

In order to evaluate the performance of our system, we used the *accuracy*, which can be computed as follows:

$$Accuracy = \frac{\# \textit{Correct}}{\# \textit{Total}}$$

where *Correct* indicates the number of correctly determined editorials, and *Total* indicates the number of total editorials.

Author	Abbr.	Train	Dev	Test	Avg. Size	Thematic Area
Abbas Güçlü	AG	355	39	37	280	Education
Attila Gökçe	AT	84	4	3	304	Sports
Derya Sazak	DS	138	32	38	403	Politics
Duygu Asena	DA	53	8	7	456	Feminine
Fikret Bila	FB	128	19	19	340	Politics
Hasan Pular	HP	257	30	42	356	General
Melih Aşık	MA	409	43	36	543	General
Sami Kohen	SK	211	35	33	391	Foreign
Taha Akyol	TA	252	44	40	320	Politics
TOTAL		1887	254	255	384	

Table 3: Number of editorials for each author of the training, development, and test data. *Avg. Size* indicates the average number of words in the editorial of that author, excluding the stopwords. The abbreviations of the authors are used in other tables

Model	Dev	Test
Word-based	91.34%	90.20%
Stem-based	83.86%	85.10%
Noun-based	80.31%	72.55%
Word-based (tuned)	92.52%	90.98%

Table 4: The results of the author detection system using word-based, noun-based, stem-based, and tuned word-based models.

4.3 Results

Table 4 presents our results using word-based, stem-based, noun-based approaches. Using the baseline word-based approach, we have reached an accuracy of more than 90% for both development and test data. Stemming, or using only nouns did not help us to improve this performance, instead we have lost 5% to 18% points using these models. So, we have returned back to the baseline word-based model, and tried to improve this. We have managed to gain one more point accuracy by optimizing the unknown word probabilities for the authors separately, instead of using only one parameter. We call this model as *tuned word-based model*.

	Guess									
Actual	AG	AT	DS	DA	FB	HP	MA	SK	TA	Accuracy
AG	39	0	0	0	0	0	0	0	0	100%
AT	0	4	0	0	0	0	0	0	0	100%
DS	1	0	29	0	0	0	0	2	0	91%
DA	0	0	0	8	0	0	0	0	0	100%
FB	0	0	0	0	15	2	0	2	0	79%
HP	0	0	0	0	3	26	1	0	0	87%
MA	0	0	1	0	1	2	39	0	0	81%
SK	0	0	1	0	0	0	0	33	1	94%
TA	0	0	0	0	0	0	0	2	42	96%

Table 5: Confusion matrix for the development set.

4.4 Error Analysis

The confusion matrix derived from the application of our approach to the development set is given Table 5. Recalling that, our system misclassified 19 of the 254 editorials in that set, 12 of these errors (about 63%) corresponds to three authors, namely, Fikret Bila, Hasan Pulur, and Melih Aşık. Note that in a similar task for Greek authors, Stamatatos *et al.* [1999] reports that 65% of the errors correspond to 3 of the 10 authors. They claim that these are the authors with an average text size smaller than others, thus unsuitable for the representation of the stylistic characteristics. My intuition is that 2 of these 3 authors write in almost all areas, not just sports, or education. This is why it is very hard to detect or differentiate them. Moreover, Hasan Pulur, when writing about politics, is confused with Fikret Bila, as seen from the confusion matrix.

5 Conclusions

We have presented a probabilistic model for automatically determining the author of a newspaper editorial. We have tried three different approaches to model the authors. First we have tried a baseline model, using only the surface forms of the words, and reached 90.8% accuracy. Then we have modeled the stems of the words, and performed worse. Finally we modeled only the stems of the nouns, and the accuracy dropped even more.

These results are important in the sense that, this work is one of the preliminary approaches in the application of corpus-based statistical methods to Turkish text processing. Mainly due to the agglutinative nature of Turkish words and the structure of

Turkish sentences, statistical methods have been largely ignored for processing Turkish. On the other hand, these characteristics of Turkish enabled us to model the authors in terms of both their word usages, by counting the word frequencies, and their stylistic preferences, by implicitly modeling their usage of suffixes. Indeed, we have not expected this interesting result, in the beginning of this work, and we assumed that, similar to morphological disambiguation, topic or sentence segmentation, name tagging, or other tasks for Turkish [Tür *et al.*, To appear; Tür *et al.*, 2000; Tür, 2000 forthcoming; Hakkani-Tür, 2000 forthcoming, among others], we could obtain some win using stemming. When we analyzed the results, we saw that, one of the main components of author attribution task is modeling stylistic features, mentioned in section 2, and indeed, Turkish words inherently contain stylistic markers in terms of suffixes. For example, the ex-prime minister of Turkey prefers a first person plural narrative style with copula in her verb usage, such as the word *gelmişizdir* (literally “We have come”), which can be morphologically analyzed as follows:

gel + Verb + Pos + Narr + A1pl + Cop

When we attempt to stem this word to the root verb *gel*, we have lost the very important stylistic information. In this respect, the agglutinative nature of Turkish is no longer a problem for this task, on the contrary, it is an advantage for this task. This is why, we have obtained worse results using the stems of the words.

One advantage of this system is that, each author gets some score of being the writer of that editorial. This enables us to combine this system with other systems, in order to get better results.

References

- [Hakkani-Tür, 2000 forthcoming] Hakkani-Tür, Dilek Z. oming. *Statistical Language Modeling for Turkish*. Ph.D. Dissertation, Department of Computer Engineering, Bilkent University, Ankara, Turkey.
- [Hankamer, 1989] Hankamer, Jorge 1989. Lexical Representation and Process. In Marslen-Wilson, W., editor 1989, *Morphological Parsing and the Lexicon*. The MIT Press.
- [Oakes, 1998] Oakes, M. P. 1998. *Statistics for Corpus Linguistics*. Edinburgh University Press. chapter 5.
- [Oflazer, 1993] Oflazer, Kemal 1993. Two-level Description of Turkish Morphology. *Literary and Linguistic Computing* 8(3).

- [Stamatatos *et al.*, 1999] Stamatatos, E.; Fakotakis, N.; and Kokkinakis, G. 1999. Automatic Authorship Attribution. In *Proceedings of EACL'99, the Ninth Conference of the European Chapter of the Association for Computational Linguistics*.
- [Stolcke, 1999] Stolcke, Andreas 1999. SRILM—the SRI language modeling toolkit. <http://www.speech.sri.com/projects/srilm/>.
- [Tür *et al.*, 2000] Tür, Gökhan; Hakkani-Tür, Dilek Zeynep; and Oflazer, Kemal 2000. Name Tagging Using Lexical, Contextual, and Morphological Information. In *Proceedings of the Workshop on Information Extraction meets Corpus Linguistics at LREC-2000, 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.
- [Tür *et al.*, To appear] Tür, Gökhan; Hakkani-Tür, Dilek; Stolcke, Andreas; and Shriberg, Elizabeth ppear. Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation. *Computational Linguistics*.
- [Tür, 2000 forthcoming] Tür, Gökhan oming. *A Statistical Information Extraction System for Turkish*. Ph.D. Dissertation, Department of Computer Engineering, Bilkent University, Ankara, Turkey.