TAR System Description Paper Template

Ivan Grubišić, Milan Pavlović, Author3

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
autor1@xxx.hr, {autor2,autor3}@zz.com

Abstract

This document provides the instructions on formatting the TAR system description paper in LaTeX. This is where you write the abstract (i.e., summary) of the work you carried out within the project. The abstract is a paragraph of text ranging between 70 and 150 words. This document provides the instructions on formatting the TAR system description paper in LaTeX. This is where you write the abstract (i.e., summary) of the work you carried out within the project. The abstract is a paragraph of text ranging between 70 and 150 words.

1. Introduction

The simplest variant of authorship attribution problem consists of determinig a true author for a given document of unknown authorship, where decision is based on a set of other documents whose authors are known (Stein et al., 2011; Ding et al., 2016). Such descirbed problem can be tackeled with supervised machine learning techniques as a single-label multiclass text classification problem, where one class represents one author (Stamatatos, 2009).

Authorship attribution problem is also known as authorship identification and it is a part of authorship analysis (Stamatatos, 2009; Ding et al., 2016). Authorship analysis is a field of stylometry which studies information about the authorship of a document, based on features derived from that document (Layton et al., 2013).

In this paper we will focus on the author diarization task proposed on PAN 2016 competition[1]. The aim of this task is to decompose a document into its authorial parts, i.e. to split a text into segments and assign an author to every segment (Koppel et al., 2011; Aldebei et al., 2015). This is one of the unsupervised variants of authorship attribution problem since text samples of known authorship are not available (Rosso et al., 2016). As we will describe, in two out of three subtasks of this task only a correct number of authors for a given document is known.

Rosso et al. (2016) divided PAN 2016 author diarization task into three subtasks. First subtask is traditionally called intrinsic plagiarism detection. The goal of this task is to find plagiariarized parts of a document in which 70% of text is written by main author and the rest by one or more other authors. The term *intrinsic* means that a decision whether plagiarized parts exist or not has to be made only by analysing a given document, without any comparisons with external sources. In the rest of the paper we refer to this subtask as a task *a*.

Other two subtasks are more related to the general task of author diarization. In the second subtask we need to segment a given document and group identified segments by author. In the rest of the paper we refer to the second subtask as a task *b*. Third subtask differs from the second one in the fact that exact number of authors is unkown. In the rest of the paper we refer to the third subtask as a task *c*.

Table 1: Basic characteristics of train datasets

Task	Number of documents	Average length (in tokens)	(min, max) authors
Task *a*	71	1679	(2, 2)
Task *b*	55	3767	(2, 10)
Task *c*	54	3298	(2, 10)

For all tasks a different training datasets are publicly available [footnote]. Rosso et al. (2016) explain that they are collections of various documents which are part of Webis-TRC-12 dataset (Potthast et al., 2013). Every document in that dataset is constructed from texts of various search results (i.e. authors) for one of the 150 topics in total. By varying different parameters such as the number and proportion of the authors, places in a document where an author switch occurs (between words, sentences or paragraphs), three training and test datasets were generated (Rosso et al., 2016). Test datasets are currently not publicly available and we could not use them for evaluation of our approach. Some basic characterisics of training datasets are shown in table 1.

- writing style, stylistic segmentation, multi-authored work, plagiarism - a, b, c

2. Related work

3. Author Diarization?

Let D be the domain of documents. We define a document $d \in D$ as a finite sequence of tokens t_i, i.e. $d = (t_1, \ldots, t_n)$ and n can differ among documents.

For each document, there is a corresponding mapping to a sequence of labels $(a_1, \ldots, a_n)$, that are representing groupings of tokens by authors. The labels a_i are indices of authors of the document, and they range from 1 to m, where m is the number of authors of the document. Each token $t \in d$ is assigned a document-level label associating it to an author. Equivalently, the codomain of the mapping can also be defined as a set of segments S. A segment s is a set of tokens where each is assigned the same label and the

[1] http://pan.webis.de/clef16/pan16-web/author-identification.html

Table 2: This is the caption of the table. Table captions should be placed *above* the table.

Model	R	P	F_1
Dummy	0	1	2
One	0	1	2
One	0	1	2
One	0	1	2

following applies:

$$t_{i_1} \in s \ \wedge t_{i_2} \in s \wedge i_1 < i_3 < i_2 \Rightarrow t_{i_3} \in s. \tag{1}$$

Also, for a set of segments to be valid, the segments must cover all terms in the document and not overlap:

$$\bigcup_{s \in S} s = T_d \wedge \bigcap_{s \in S} s = \{\}, \tag{2}$$

where T_d is the set of all terms in document d.

Let $\mathcal{D}$ be a

- a
- b

features

differences: fixed features + transformation vs document-dependent features

4. Experimental results

1 baselines

Mention cofidences.

4.1. Intrinsic plagiarism detection

setup results

4.2. Author diarization with known numbers of authors

setup

4.3. Author diarization with unknown numbers of authors

setup

5. Conclusion

Conclusion is the last enumerated section of the paper. It should not exceed half of a column and is typically split into 2–3 paragraphs. No new information should be presented in the conclusion; this section only summarizes and concludes the paper.

6. Further work

Acknowledgements

If suitable, you can include the *Acknowledgements* section before inserting the literature references in order to thank those who helped you in any way to deliver the paper, but are not co-authors of the paper.

References

Khaled Aldebei, Xiangjian He, and Jie Yang. 2015. Unsupervised Decomposition of a Multi-Author Document Based on Naive-Bayesian Model. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 501–505.

Steven HH Ding, Benjamin Fung, Farkhund Iqbal, and William K Cheung. 2016. Learning stylometric representations for authorship analysis. *arXiv preprint arXiv:1606.01219*.

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1356–1364. Association for Computational Linguistics.

Robert Layton, Paul Watters, and Richard Dazeley. 2013. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*, 19(01):95–120.

Martin Potthast, Matthias Hagen, Michael Völske, and Benno Stein. 2013. Crowdsourcing interaction logs to understand text reuse from the web. In *ACL (1)*, pages 1212–1221.

Paolo Rosso, Francisco Rangel, Martin Potthast, Efstathios Stamatatos, Michael Tschuggnall, and Benno Stein. 2016. Overview of pan'16. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 332–350. Springer.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Benno Stein, Nedim Lipka, and Peter Prettenhofer. 2011. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82.