Information flow between news articles: Slovene media case study

Jan Chołoniewski
Center of Excellence for
Complex Systems Research,
Faculty of Physics,
Warsaw University of
Technology,
Koszykowa 75, PL-00662,
Warsaw, Poland
choloniewski@if.pw.edu.pl

Gregor Leban
Artificial Intelligence
Laboratory,
Jožef Stefan Institute,
Jamova 39, 1000 Ljubljana,
Slovenia
gregor.leban@ijs.si

Sebastijan Maček,
Aljoša Rehar
Slovenska Tiskovna Agencija,
Tivolska 48, 1000 Ljubljana,
Slovenia
{sm,ar}@sta.si

ABSTRACT

We present results of a study on usage of text similarity measures based on co-occurrence of words and phrases to classify a relation between a pair of news articles (i.e. no relation, both based on a common source, one based on the other). For each Slovenian article written in Slovene and published online on 27th June 2016, we found the most similar release from the Slovenian Press Agency (STA) database to obtain a list of candidate article-source pairs. Four experts from STA were asked to score the pairs, and their annotations were used to train classifiers and evaluate their accuracy.

1. INTRODUCTION

Propagating, exchanging, organizing and processing information are important parts of human social interactions on both micro- [1] and macro-level [2]. After years of local and nationwide scale, newspapers, press agencies and news outlets started to operate at global level using Internet. An easy and open access to their releases is desirable for news consumers and can be tracked e.g. with website traffic statistics or in social media. Article reuse by other publishers (authorized or not) is however not that straight-forward.

Combining natural language processing methods with data gathered in the Internet allows to quantify and measure social information processing phenomena [3, 4, 5, 6]. The advancements in NLP might serve all types of text-based media (in particular online news outlets) to provide tools to track spreading of their texts.

A tool that automatically finds articles based on a given article might be useful for news outlets and press agencies to track usage of their releases and to find cases of plagiarism or unauthorized use. Moreover, it might be applied to large scale news spreading studies [3]. A software-assisted plagiarism detection is a well-known problem in an information retrieval field [7], and using text similarity-based methods is one of the most popular approaches [8]. To the best of our knowledge, the following paper is the first published study of plagiarism detection in Slovene media supported by professional press agency workers.

The aim of the presented work is to check if text comparison methods based on co-occurrence of phrases can be successfully applied to determine a relation between two articles. Possible relations that we want to determine are (a) there is no relation, (b) they share a common source, or (c) one is based on the other one. To find the most efficient way to do that, we calculated cosine similarity of "bag of n-grams" representations of articles from Slovene media published on one day with releases from Slovenska Tiskovna Agencija (STA; Slovenian Press Agency) to preselect the most similar release to each article, asked experts to annotate the candidate pairs, and compared results for different thresholds and n-grams with the annotations.

The rest of the paper is structured as follows: in Section 2 we highlight a process of obtaining experimental data (candidate source release matching, expert annotation study), in Section 3 we describe applied methods and benchmark parameters, then in Section 4 we present results of classification study in two simplified cases; Section 5 contains a discussion and possible improvements, and Section 6 sums up the research.

2. DATA

Data for the study consist of randomly selected 469 articles out of 895 published on 27th June 2016 from 62 Slovene online news outlets as tracked by the EventRegistry [9]. For each article, we have found the most similar one in the STA releases database in terms of cosine similarity of two-, three- and four-word phrases ({2,3,4}-grams) occurrence vectors with TF-IDF weighting (see section 3 for details). A histogram of obtained similarities is presented in Figure 1. About 75% (354 out of 469) of candidate pairs obtained a cosine similarity below 0.1, and 10% (47 out of 469) over 0.9.

The pairs were scored by four experts from STA (A1, A2, A3, A4). They were asked to mark each pair with one of the following scores:

- NF – the proper source release has not been found despite it is present in the STA database,
- NC – the proper source release has not been found and it is not present in the STA database,
- DS – the proper source release has been found (although it might be one of many sources of the article),

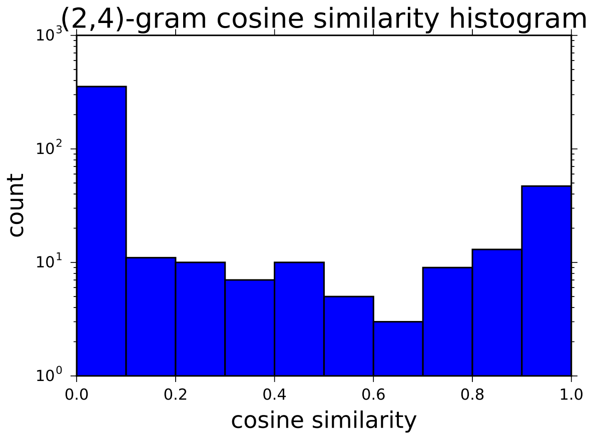

Figure 1: A histogram of {2,3,4}-gram cosine similarities of candidate pairs with a logarithmic Y-axis.

- IDS – the article and the proposed source release are both based on the same third party source.

In cases where the source was not found (NF), the annotators provided a link to the proper source release.

In Table 1, we present basic statistics of the annotations given by experts. We considered two methods of simplifying the annotations. The first one (A), merges DS and IDS marks to discriminate between two classes – a given pair contains pieces of the same information or is unrelated. The second one (B), merges IDS with NC – the algorithm's task is to check if one text is directly based on the other one.

person	total	NF	NC	DS	IDS
A1	469	3	315	98	53
A2	469	2	358	97	12
A3	95	0	61	23	11
A4	95	0	70	20	5

Table 1: Basic statistics of raw candidate release-article pairs annotations by the STA experts. total – a number of annotated pairs; NF – source not detected despite the source release is in the STA archive; NC – no source release in the STA archive; DS – one article is a direct source of the other; IDS – both documents based on the same third source article.

In Table 2, percentages of agreement among annotators are being presented for (a) raw annotations, (b) simplification A and (c) simplification B (see above).

The annotators were sometimes non-unanimous when both articles in a pair had a common source (compare Table 2a and 2b, mean agreement = 87%). They were more consistent when a release was a source of a given article (compare Table 2a and 2c, mean agreement = 96%).

Additionally, because of score inconsistencies, the final list has been prepared after discussing problematic cases.

	A1	A2	A3	A4
A1	100%	87%	86%	87%
A2	87%	100%	89%	89%
A3	86%	89%	100%	83%
A4	87%	89%	83%	100%

(a) Raw annotations

	A1	A2	A3	A4
A1	100%	88%	86%	88%
A2	88%	100%	91%	89%
A3	86%	91%	100%	84%
A4	88%	89%	84%	100%

(b) Simplified A – DS and IDS merged

	A1	A2	A3	A4
A1	100%	96%	99%	95%
A2	96%	100%	99%	96%
A3	99%	99%	100%	95%
A4	95%	96%	95%	100%

(c) Simplified B – IDS and NC merged

Table 2: Agreement among annotators.

3. METHODS

Articles and releases were mapped to "bag of n-grams" representations. Additionally, n-gram counts were transformed using term frequency-inverted document frequency (TF-IDF) weighting trained on a corpus of 5,000 randomly selected Slovene articles stored in the EventRegistry published during two weeks preceding the analyzed day. Terms which occurred in more than 25% of documents were discarded. Laplace smoothing with $\alpha = 1$ was applied to include terms which were not present in the corpus.

For $n = 1, ..., 5$, weighed term vectors of Slovene articles from 27th June 2016 were compared with all vectors of STA releases published between 20th and 27th June 2016 to find candidate source releases. For each n, we tested classifiers with a threshold from 0.00 to 1.00 with steps of 0.01 to find the threshold for which the method achieves the highest F1-score for A and B simplifications separately. The releases were compared with the list created using preselected pairs and experts' comments. A source release for a given article and a given threshold was considered as correctly found, when it matched the one annotated by human and cosine similarity score was above the threshold. A given article was considered correctly classified if a source release was correctly found or if the article was correctly marked as not having a source release in the STA database.

Parameters used to score the classification were accuracy, recall, precision, and F1-score. We used following definitions:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FN}$$

$$F1 = 2\frac{recall \times precision}{recall + precision}$$

where TP – number of articles with a correctly found source, TN – number of articles correctly marked as not having source in the STA database, FP – number of articles incorrectly marked as having source in the database, FN – number of articles incorrectly marked as not having source in the database. Cases when articles had incorrectly found source were counted separately as *errors*.

Each annotator could have scored differently each article-source pair thus the mean values and standard deviations of parameters were calculated when considering lists of annotations separately.

4. RESULTS

For each n value, we have found a threshold which maximized mean F1-score over all annotators. Results are shown in Table 3a for the simplification A and in Table 3b for the simplification B.

n	threshold	acc	σ_{acc}	F1	σ_{F1}	errors
1	0.29	0.90	0.02	0.83	0.04	18
2	0.09	0.91	0.02	0.84	0.03	4
{2,3,4}	0.06	0.91	0.01	0.84	0.02	3
3	0.05	0.91	0.01	0.84	0.02	4
4	0.04	0.90	0.02	0.83	0.02	3
5	0.03	0.90	0.02	0.83	0.02	6

(a) Simplified A – direct and indirect relations merged

n	threshold	acc	σ_{acc}	F1	σ_{F1}	errors
1	0.56	0.95	0.01	0.88	0.03	4
2	0.46	0.96	0.01	0.90	0.04	1
{2,3,4}	0.27	0.96	0.01	0.90	0.03	1
3	0.25	0.96	0.01	0.90	0.03	1
4	0.22	0.96	0.01	0.90	0.03	1
5	0.13	0.95	0.01	0.89	0.02	2

(b) Simplified B – indirect relations and lacks of relation merged

Table 3: Thresholds resulting with the best F1 for different ns. acc – mean accuracy, σ_{acc} – standard deviation of accuracy, F1 – mean F1-score, σ_{F1} – standard deviation of F1-score, errors – mean number of incorrectly found sources.

The results for the simplification A are satisfying when compared to the agreement among annotators. There were no significant difference between specific $n > 1$ but for $n = 1$ there were as many as 18 errors. The results for the simplification B are comparable with an agreement among annotators (see Table 2) and the classifiers could not find a correct source only in one case in which article was mainly based on some other release and only partially on the detected one. Again, there is very little difference among different n-grams which might suggest that in most cases articles use similar phrasing as the source release and the method is efficient.

In Figures 2 and 3, we show a histogram of cosine similarities and a stacked bar plot showing fraction of each score in the final list in each cosine similarity bin for $n = 1$ and $n = 3$ (respectively). The cosine similarities are not dramatically more separated in any of the cases but using $n = 1$ leads to significantly higher number of errors, and using $n = 5$ – to a slight increase of number of errors.

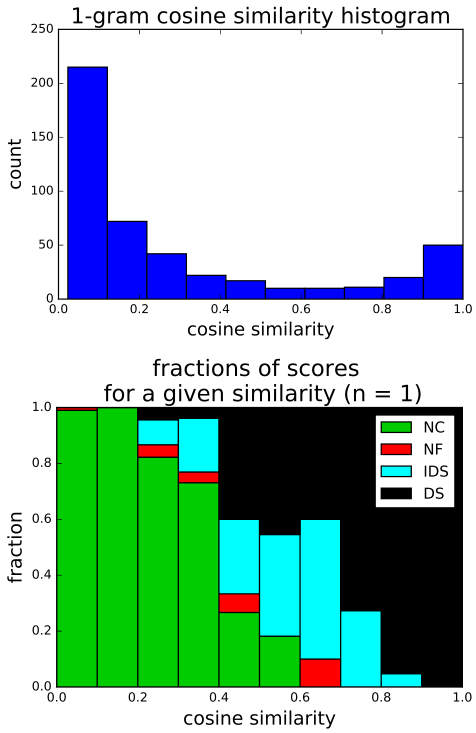

Figure 2: (top) A histogram of n-gram cosine similarities and (bottom) a fraction of each score in each similarity bin (see Section 2 for abbreviation expansions) for $n = 1$.

5. DISCUSSION

For most values of n, over 85% of candidate pairs had extreme cosine similarity values (below 0.1 or over 0.9). Two articles with cosine similarity equal to 1 are duplicates while the articles with cosine similarity equal to 0 are completely unrelated. Similarities between those values are not that clear to interpret. Obtaining more pairs with intermediate values would make results for boundary cases more reliable. After closer examination, very similar pairs which were marked as unrelated turned out to be annotators' mistakes. On the other hand, in the opposite cases (pairs with low cosine similarity but marked as related) the analyzed articles were rewritten; using lemmatization might be sufficient to identify them as similar.

Using different ns did not cause significant changes of accuracies and F1-scores of classifiers in both simplified cases but $n > 1$ allows to correctly find more sources than $n = 1$. In most $n = 1$ errors, the algorithm pointed at some more general release about a given topic.

We considered three types of relations between text pairs – lack of relation, common source, and direct sourcing (one based on the other). For the first and the last types of relation, it was usually possible to distinguish between them

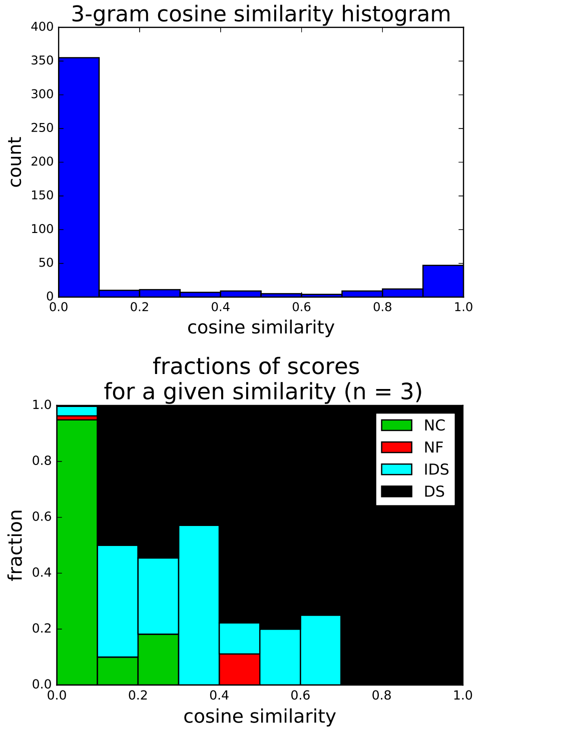

Figure 3: (top) A histogram of cosine similarities and (bottom) a fraction of each score in each similarity bin (see Section 2 for abbreviation expansions) for $n = 3$.

but in the proposed way it was not possible to accurately identify when two articles had a common source.

The method we used has not created a completely clear separation between considered relation types. In the further work the approach could be improved with lemmatization, mapping to WordNet synsets, discarding proper nouns, or a proper treatment of quotations.

It is also important to take into account that the experts were able to discriminate pairs because of their domain-specific knowledge. Nevertheless, even highly trained individuals scored some pairs differently. In many cases, there can be more then one source release of an article or an article might be based only partially on a given release.

An important future work will include use of cross-lingual techniques (e.g. [10]) to compute similarities and detect plagiarism in news articles in different languages.

6. CONCLUSIONS

We have presented a case study of estimating usage of STA releases by Slovene news outlets. We applied ”bag on n-grams” representations of articles and releases with TF-IDF weighting, and compared them pairwise using cosine similarity. Detected candidate ”article-source release” pairs were annotated by experts.

We compared results of automatic source detection with the annotations, and as expected found that articles have higher cosine similarity to releases when they are directly based on them, and can be detected with about 96% accuracy. A discrimination between not related and related pairs was possible with a 90% accuracy.

The results might be useful for a broader use although a partial supervision in boundary cases would be required. We suspect that lemmatization, proper quotations filtering and discarding proper nouns might result in achieving higher accuracies. Using cross-lingual similarity measures would be another interesting modification.

7. ACKNOWLEDGMENTS

We thank Aleš Pečnik for a technical support, and Tjaša Doljak, Naum Dretnik and Sabina Zonta from STA for annotations. This research has received funding as *RENOIR* Project from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 691152. JCh has been also supported by Ministry of Science and Higher Education (Poland), grant No. 34/H2020/2016.

8. REFERENCES

[1] M.N. Bechtoldt, C.K.W. De Dreu, B.A. Nijstad, and H.-S. Choi. Motivated information processing, social tuning, and group creativity. *J. Pers. Soc. Psychol.*, 99(4):622–637, 2010.

[2] O. Oh, M. Agrawal, and H.R. Rao. Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quart.*, 37(2):407–426, 2013.

[3] K. Lerman. Social information processing in news aggregation. *IEEE Internet Comput.*, 11(6):16–28, 2007.

[4] V. Niculae, C. Suen, J. Zhang, C. Danescu-Niculescu-Mizil, and J. Leskovec. QUOTUS: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*, WWW ’15 Companion, pages 798–808, 2015.

[5] A. Chmiel, J. Sienkiewicz, M. Thelwall, G. Paltoglou, K. Buckley, A. Kappas, and J.A. Hołyst. Collective emotions online and their influence on community life. *PLoS ONE*, 6(7), 2011.

[6] J. Chołoniewski, J. Sienkiewicz, J. Hołyst, and M. Thelwall. The role of emotional variables in the classification and prediction of collective social dynamics. *Acta. Phys. Pol. A*, 127(3):A21–A28, 2015.

[7] A Parker and JO Hamblen. Computer algorithms for plagiarism detection. *IEEE T. Educ.*, 32(2):94–99, 1989.

[8] D. Metzler, Y. Bernstein, W.B. Croft, A. Moffat, and J. Zobel. Similarity measures for tracking information flow. In *Proceedings of International Conference on Information and Knowledge Management*, pages 517–524, 2005.

[9] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. Event registry: Learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW ’14 Companion, pages 107–110, 2014.

[10] http://xling.ijs.si.