Analogical News Angles from Text Similarity

Bjørnar Tessem

Department of Information Science and Media Studies
University of Bergen, Norway
bjornar.tessem@uib.no

Abstract The paper presents an algorithm providing creativity support to journalists. It suggests analogical transfer of news angles from reports written about different events than the one the journalist is working on. The problem is formulated as a matching problem, where news reports with similar wordings from two events are matched, and unmatched reports from previous cases are selected as candidates for a news angle transfer. The approach is based on document similarity measures for matching and selection of transferable candidates. The algorithm has been tested on a small data set and show that the concept may be viable, but needs more exploration and evaluation in journalistic practice.

Keywords: Computational creativity, Analogical reasoning, Document similarity, Journalism

1 Introduction

Artificial intelligence is considered to have great potential in journalism [10], already found in robot journalism [12], content verification [6], and data analysis [8]. One way to go is to support the journalist creativity by providing suggestions for new angles to a new report on an event, e.g., a new news paper article. This is the aim of the News Angler project where we aim to support journalists with such creativity tools [5,11].

The term *news angle* was coined already in the seventies by Altheide [1] who observed that reporters rely on "'angles,' or story lines, which give the specific events new meaning". So, finding a new angle on an event is what the reporter relies on to make the report interesting for a user even though the event already has been described in several reports and in many news media.

One approach to proposing news angles to the journalist is to find and suggest reports from other events that are similar to the current event, but with angles that have not been tried on the current event. This constitutes a form of analogical reasoning where an algorithm identifies unmatched aspects of a base case and transfers them to a new target case, parallel to the *transfer* part of Falkenhainer et al.'s structure mapping engine [4]. This paper describes an analogical search algorithm that uses text similarity metrics for news reports and events to identify reports that can provide the journalist with an unused news angle. Even with a simple technique like the use of tf-idf (term frequency - inverse document frequency [9]) we are able to see some promising results.

2 Assumptions

Any news event consists of entities, most often humans, their properties, relations, situations and sub-events that transform the state of some entity [11]. It is the journalist's task to pick a subset of these features from an event and present them in a report, and it is this subset that can be considered the particular angle on the event. Here, these entities, properties, relations, states, and situations are not explicitly known, but are *externalised* in news reports that we use.

Thus, the collection of journalists that have reported from the event is seen as data generating entities. In each report they tell about the event using one or a few angles. Further, in the reports the choice of angle(s) will influence the final wording. The final wording may then be used to compute text based similarities among reports. finally, events themselves have the collected set of reports and additional information about the entities from other sources (e.g. Wikipedia) as an input to a general event similarity, which may also be computed from text similarity metrics.

3 Finding Unmatched News Reports

An optimization approach is used to identify unmatched reports in an analogical event. Assume that we have a target event τ with n_τ news reports $t_j \in T$ that we want to find a new angle for, as well as an *identified and similar* base event β that has n_β base news reports $b_i \in B$. Also assume that we have a similarity measure $sim(b_i, t_j) \in [0, 1]$ for each pair of reports $b_i \in B$ and $t_j \in T$. See Section 4 for realisations of similarity measures.

Now, let A be a binary matrix with entries $a_{ij} = 1$ if there is a *match* between $b_i \in B$ and $t_j \in T$, otherwise 0. A represents the total matching between base and target. The idea is that a matching between reports indicates that they have similar or same angles. There is a couple of domain based heuristic constraints, in addition to maximum similarity among matched reports, that should be fulfilled for a matching to have high quality. First, reports with low similarity should not be matched; second, reports should usually not match more than one other report.

To handle the problem with low similarity we may subtract a constant c_l from all similarity values to ensure that all matched reports have a similarity above the limit c_l. To ensure almost one-to-one matching we introduce a penalty for having more than one match in a row or column. So we need to count the number of 1's in each row (cr_i) and each column (cc_j) of A. The penalty for having more than one 1 in a row or column is c_p. A matching of high quality is then found by maximizing the objective function

$$f(A) = \sum_{i,j} a_{ij}(sim(b_i, t_j) - c_l) - \sum_i max(0, cr_i - 1) \cdot c_p - \sum_j max(0, cc_j - 1) \cdot c_p$$

The matching A can be found in a greedy manner by maintaining a sorted list L of indices (i, j) referring to reports $b_i \in B$ and $t_j \in T$ that may be matched.

We include only the pairs with a positive $sim(b_i, t_j) - c_l$ in L, as the others will contribute negatively to the total matching score. For each index pair we also maintain a $gain(i, j) = -sim(b_i, t_j) + c_p \cdot (ind(i) + ind(j))$ where $ind(i) = 1$ if i is found in more than one candidate pair in L, otherwise 0 (similar for $ind(j)$). We repeatedly remove the pair with most gain from L, and update the gain for the remaining pairs. When there are no pairs with positive gains left, L represents an optimal matching A, where a_{ij} is 1 if L contains the pair (i, j), 0 otherwise.

When we have found the solution A, there will be reports about the base event which are unmatched, i.e., there are rows in A where all entries are 0. Each of these unmatched base reports may suggest a new angle. Journalists could be responsible for investigating the candidates, but may need some guidance. The most relevant candidate could for instance be the unmatched report that has the highest similarity to any existing report in the target, i.e., has the highest $rel_\tau(b_i) = \max_j sim(b_i, t_j)$.

From here, it is possible to rank candidates from all possible base events β_k by combining event similarity with the relevance score for each report. For now, let us assume that we are able to compute the event similarity $sim(\beta_k, \tau)$ for all base events β_k and the target τ (See section 4). Further, assume that we for each β_k have an optimal report matching A^k. All unmatched reports in the events β_k will now be candidates for a transferred angle. To rank all these selected reports, we use the event similarities as well as the relevance-measure $rel_\tau(b_i)$:

$$score_\tau(b_i^k) = sim(\beta_k, \tau) \cdot rel_\tau(b_i^k)$$

4 Similarity Measures

There are many ways of measuring text similarity; this includes the use of standard IR techniques like tf-idf[9], the use of topic modeling[2], word2vec[7], graph2vec[13] (provided we are able to lift the knowledge about the event and its reports into knowledge graphs), and most recently the BERT[3] and XLNet[14] frameworks. The outcome of the analogical search algorithm presented above will depend on the quality of the similarity measures we use, so there is a need to experiment with these.

The tf-idf model of document similarity is a natural starting point and will serve as a base line for further explorations of the general algorithm. So far we have been able to run tests on a small collection of ten events with 20 reports each, with Wikipedia articles (about 20 in each event) about entities occurring in the events as supporting data. To run tf-idf models we have relied on the Python gensim library for text processing[1]. All texts were lemmatized using gensim algorithms and only verbs, nouns, adjectives, and adverbs were included.

The eleven tf-idf models in use were:

- one for the whole collection of events, where each event's reports and Wikipedia texts were concatenated into one text document. This gave us a doc-

[1] https://radimrehurek.com/gensim/

ument base of 10 large documents, enabling us to get a similarity measure for each pair of events.
- one for each of the ten events, where the document collection was the individual reports and the Wikipedia articles. These models allow us to compute similarities between any report and each of the reports of the event, for example $sim(b_i, t_j)$. Thus, similarity to reports of a particular event is based on the reports of that event itself only.

5 Results

The data for these initial experiments were reports from 10 events collected in March 2019. The events (and two letter codes for later references) are

College scandal (CS) Wealthy Americans getting their children into prestigious schools by paying school officers.

Zuma nepotism (ZN) Previous South African president awards political positions to rich people who supports his family economically.

Barry Bonds case (BB) Disclosure of doping tests that showed that famous baseball player Barry Bonds were doped in parts of his career.

Penelopegate (PG) French president candidate used his position to give family members public positions.

Menendez corruption (MC) Democratic senator accused of accepting gifts from wealthy friend in exchange of favors in political decisions.

Armstrong doping (AD) The doping case against world famous cyclist Lance Armstrong.

Sudan protests (SP) Series of demonstrations against long term Sudanese president Omar al-Bashir.

Russian doping (RD) Systematic government supported doping in Russian sport.

Trudeau scandal (TS) Politician close to Canadian prime minister Justin Trudeau illegally influenced the justice system on behalf of a Canadian construction company.

Mueller report (MR) The release of the Mueller report about Russian meddling with the 2016 presidential election in USA.

In the experiments, most computed similarities between events were small (less than 0.01). Anyhow, here it is the relative sizes that count, as a ranking is more interesting than the numbers themselves. However, notice that the three doping events have the highest similarities $sim(\mathrm{BB}, \mathrm{AD}) = 0.079$, $sim(\mathrm{BB}, \mathrm{RD}) = 0.103$, and $sim(\mathrm{AD}, \mathrm{RD}) = 0.270$ indicating that wording in the reports on these three cases are very similar, and containing specific doping related words.

The next step was to compute for each event (as a target event) the potential unmatched reports from each of the other events (as base events). We used the matching algorithm, calculated relevance scores $rel_\tau(b_i^k)$ and further a total $score_\tau(b_i^k)$ for all unmatched reports. Results showing the most promising transfer candidate for each target event are found in Table 1. The title of the report with most promising new angle is given for each event, and also a suggestion for a journalistic transfer of the angle.

Table 1. Suggested transfers of angles

Event	Article title for transfer	Journalistic angle
CS	BB: Lawyer jailed for leaking steroids testimony	Has anyone been convicted?
ZN	SP: Sudan protesters move to protect Khartoum	No immediate angle
BB	RD: Russian Olympic team's drug usage could have long term effects on athletes' health	Has Barry Bond's health been influenced by doping?
PG	AD: Cycling bosses slammed over Lance Armstrong	What do powerful people think of Penelopegate?
MC	MR: Barr scours Trump-Russia report to see how much to open	No immediate angle
AD	RD: Russian doping said to run deep	Are there powerful people involved in Armstrong's doping?
SP	ZN: Zuma plea as protests sweep the townships: South Africa's president calls for an end to the violence as he admits that he needs time to end corruption and improve government services	What does al-Bashir say to protesters?
RD	AD: Armstrong's biggest sponsors sever ties	How are sponsors of Russian sport reacting?
TS	ZN: In Gupta Brothers' Rise and Fall, the Tale of a Sullied A.N.C.	What does the scandal mean for the reputation of the Liberal Party?
MR	RD: 'My message to the British runners who lost to our drug cheats? Sorry'	Has Mueller a comment to the Democrats about the election meddling

6 Conclusion and Further Work

This paper has described initial work on a tool for providing journalists with information that may suggest a new angle to an event. Here we have presented an algorithm that suggests for a journalist working on a particular event, the transfer of news angles found in reports of a different event, based on document similarity and a form of analogical reasoning. The results so far are not much more than a proof-of-concept, but show some interesting results, even with unsophisticated methods for document similarity.

The suggestions for journalistic angles here are suggestions based on our own perceptions, and we found a plausible one for eight of the ten events. Practicing journalists may think otherwise about what angles are interesting, and the results need to be validated against their opinions, i.e., which report from base events gave the best idea for a new news angle. We need to set up experiments with journalists for this purpose. A second important task is to explore other similarity measures. The algorithm itself will be valid, but may get better results from improved document similarity measures for instance taking into account context sensitivity.

Acknowledgement. The News Angler project is funded by the Norwegian Research Council's IKTPLUSS programme as project 275872.

References

1. Altheide, D.L., Rasmussen, P.K.: Becoming news: A study of two newsrooms. Sociology of Work and Occupations **3**(2), 223–246 (May 1976)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3**(Jan), 993–1022 (2003)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Falkenhainer, B., Forbus, K.D., Gentner, D.: The structure-mapping engine: Algorithm and examples. Artif. Intell. **41**(1), 1–63 (1989)
5. Gallofré Ocaña, M., Nyre, L., Opdahl, A.L., Tessem, B., Trattner, C., Veres, C.: Towards a Big Data Platform for News Angles. In: Proceedings of the 4th Norwegian Big Data Symposium (NOBIDS 2018). vol. 2316, pp. 17–29. CEUR Workshop Proceedings (Nov 2018)
6. Gravanis, G., Vakali, A., Diamantaras, K., Karadais, P.: Behind the cues: A benchmarking study for fake news detection. Expert Systems with Applications **128**, 201–213 (Aug 2019)
7. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 957–966. PMLR, Lille, France (07–09 Jul 2015)
8. Lewis, S.C., Westlund, O.: Big data and journalism: Epistemology, expertise, economics, and ethics. Digital Journalism **3**(3), 447–466 (2015)
9. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge Univ. Press, New York (2008)
10. Miroshnichenko, A.: AI to Bypass Creativity. Will Robots Replace Journalists? (The Answer Is "Yes"). Information **9**(7), 183 (Jul 2018)
11. Opdahl, A.L., Tessem, B.: Towards Ontological Support for Journalistic Angles. In: Reinhartz-Berger, I., Zdravkovic, J., Gulden, J., Schmidt, R. (eds.) Enterprise, Business-Process and Information Systems Modeling. pp. 279–294. Lecture Notes in Business Information Processing, Springer International Publishing (2019)
12. Simonite, T.: Robot Writing Moves from Journalism to Wall Street (2015), https://www.technologyreview.com/s/533976/robot-journalist-finds-new-work-on-wall-street/
13. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: Proc. 21st AAAI, February 4-9, San Francisco, USA. pp. 4444–4451 (2017)
14. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs] (Jun 2019), http://arxiv.org/abs/1906.08237, arXiv: 1906.08237