Human-centric evaluation of similarity spaces of news articles

Clara Higuera Cabañes Michel Schammel Shirley Ka Kei Yu Ben Fields

[first name].[last name]@bbc.co.uk
The British Broadcasting Corporation
New Broadcasting House, Portland Place
London, W1A 1AA
United Kingdom

Abstract

In this paper we present a practical approach to evaluate similarity spaces of news articles, guided by human perception. This is motivated by applications that are expected by modern news audiences, most notably recommender systems. Our approach is laid out and contextualised with a brief background in human similarity measurement and perception. This is complimented with a discussion of computational methods for measuring similarity between news articles. We then go through a prototypical use of the evaluation in a practical setting before we point to future work enabled by this framework.

1 Introduction and Motivation

In a modern news organisation, there are a number of functions that depend on computational understanding of produced media. For text-based news articles this typically takes the form of lower dimensionality content-similarity. But how do we know that these similarities are reliable? On what basis can we take these computational similarity spaces to be a proxy for human judgement? In this paper we address this question as follows.

- How can we assess human cognition of the similarity for news articles
- Analogously, what are efficient and effective means of computing similarity between news articles
- By what means can we use the human cognition of article similarity to select parameters or otherwise tune a computed similarity space

A typical application that benefits from this sort of human calibrated similarity space for news articles is an article recommender system. While a classic collaborative filtering approach has been tried within the news domain [LDP10], typical user behaviour makes this approach difficult in practice. In particular, the lifespan of individual articles tends to be short and the item preferences of users is light.

This leads to a situation where in practice a collaborative filtering approach is hampered by the cold-start problem, where lack of preference data negatively impacts the predictive power of the system. To get around this issue, a variety of more domain-specific approaches have been tried [GDF13, TASJ14, KKGV18]. However, these all demand significant levels of analytical effort or otherwise present challenges when scaling to a large global news organisation. A simple way to get around these constraints while still meeting the functional requirements[1] of a recommender system is to generate a similarity space across recently published articles and be able to surface the most similar content to the current article. This assumes that most readers predominantly prefer reading similar content, but this a pragmatic assumption.

In order for this approach of article similarity to be an effective means for recommendation to readers, the similarity space needs to be well aligned with the human perception of similarity across these articles.

In: A. Aker, D. Albakour, A. Barrón-Cedeño, S. Dori-Hacohen, M. Martinez, J. Stray, S. Tippmann (eds.): Proceedings of the NewsIR'19 Workshop at SIGIR, Paris, France, 25-July-2019, published at http://ceur-ws.org

[1] Here that means: present a reader of an article with other articles that they have a high likelihood of reading

To that end, this paper will lay out a methodology for assessing the perception of similarity between news articles (Section 2), methods for computing similarity between news articles (Section 3), and an example case where findings from the first part are used to aid model selection in the second (Section 4). We also briefly discuss how such a content similarity recommender system works in practice before we conclude the paper by considering next steps implied by this work.

2 Human Similarity

Given that our motivation for having a similarity space among news articles is to produce articles that readers perceive as similar, it is critical that we have a means of assessing similarity of news articles, as perceived by people. While it would be convenient to assume that news articles are perceived by people as having objective similarities, there are a number reasons to work from the assumption that is not the case. Broadly, human perception of item similarity does not obey the requirements of a well-formed metric space, most notably symmetry [AM99] and the triangle inequality [YBDS$^+$17].

Therefore we look to other domains for useful analogues to our problem of assessing the perceptual difference between objects and a mapping of that into a similarity metric. In particular, we look at assessment methods from two domains: psychophysics and sensory perception.

2.1 Psychophysics

The field of psychophysics is concerned with understanding the interaction between physical phenomena and human cognition of these phenomena, most typically auditory and visual stimulus. One of the most widely known applications from psychophysics is lossy compression, where digital audio or video is reduced in size by discarding portions that are not likely to be perceived by a general audience[Pan95, Wal92]. As a result of these well established areas of research, this field has mature techniques for measuring human-perceivable difference across transformations or deterioration of an anchor stimuli. The standard practice in auditory settings is called Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) [15301]. This testing framework allows for the precise measuring of change which are or are not generally noticeable while calibrating for individual testers' differences in perception and cognition, though this comes at the expense of a test which can be lengthy and require larger populations of testers than less complicated tests.

2.2 Sensory Perception

A common means of measuring the human ability to differentiate between stimuli that are similar is described in terms of *Just Noticeable Difference* (JND). That is, the JND is a unit where if two stimuli are *measurably* closer than this JND, the average person will not be able to notice the difference between these stimuli. This has been effectively used to understand human perception of a wide variety of things from speech [BRN99] or colour [CL95] to the handling characteristics of cars [HJ68]. In a news article context the JND is the amount of measurable change between articles before an average reader would consider them different articles.

Serving as a complement to the idea of JND is a sensory triangle test. In this test three stimuli are presented to an evaluator, with two of them being identical. The evaluator is then asked to identify which of the three stimuli is different from the other two. This process is repeated by a population of evaluators, and if a statistically significant[2]portion of the population correctly identifies the different stimuli, the difference is taken as perceivable and therefore larger than the JND [OO85].

2.3 A Proposed Test

Given the above, we propose the following means of assessing article similarity.

1. Gather a collection of *anchor* articles from your corpus.
2. For each anchor select two additional articles for comparison
3. Present each of these triplets in turn to a human evaluator asking the evaluator to decide which of the two articles is most similar to the anchor

Beyond the evaluation process, there is the mechanism for selecting both the anchors and the comparison articles. For these issues much depends on the particulars of the assessment and to that end we will go through our use of this assessment in Section 4. However, there are some guiding principles to consider in general. Keeping in mind that the goal of the assessment is a human understanding of the similarity space, rather than the analytical configuration of the space, we should seek to select anchors to maximise coverage across the corpus and we should seek to select comparison articles that we believe to be a variety of different levels of similarity from the anchor articles. A straightforward way to bootstrap these selection criteria is to

[2]typically a chi-squared test is used, c.f. `https://www.sensorysociety.org/knowledge/sspwiki/pages/triangle\%20test.aspx`

use a best-effort computed similarity and to the select items across the space.

By adhering to these principles we should be able to improve our results, though as with many assessments of this type, the larger the number of participants becomes, the stronger the conclusion will be.

3 Computed Similarity

In order to compute a similarity measure between articles, we first need to derive a computer-readable representation for each document and second, choose an adequate metric to evaluate the *distance* between them.

There are several algorithms that can be used to construct similarity spaces and perform topic modelling.

3.1 Doc2vec

Word2vec [MCCD13] and its extension to Doc2vec [LM14] are embedding algorithms (usually formed of shallow, two-layer neural networks) that construct vector spaces of words based on their frequencies and cooccurrences in the training corpus. The hence learned mathematical representation can be used to establish similarities between words using vector algebra. Doc2Vec works in a similar way but trains on individual documents rather than words and is thus able to establish similarities between documents rather than just words.

3.2 FastText

Another popular natural language processing library is fastText. Based on a shallow neural network with an embedding layer, fastText can be used in two applications: learning embeddings from a corpus [BGJM17] or document classification [JGBM17]. In the former application, [GBG+18] used the fastText algorithm to generate language models for 157 different languages from Wikipedia data. These pre-trained models can be used to transform documents into vector representation and enable similarity calculations in the same manner as in the Doc2vec case.

3.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [BNJ03] is a generative probabilistic model that represents documents as a mixture or collection of topics expressed as probabilities with each topic represented by a probability distribution of words. Section 3.4 describes how the similarity between documents can be assessed with this method.

For our use case, we found LDA has a number of advantages:

- The algorithm delivers inspectable topics; as every topic is a probability distribution of words, it is straightforward to determine the most important words contributing to each topic and thus allowing interpretation of the topics.
- Building onto the word distributions, the topics associated with a document can easily be traced back to the most salient words in the document. This is a strong step towards explainability; a key requirement under recital 71 of the GDPR [RP16] and a strong tool for recommender monitoring.

3.4 Similarity Measures

In order to compute similarity between documents, one requires the use of a metric, which, in the case of vector spaces, usually resorts to Euclidean distance or cosine similarity. However, in the case of probability distributions, a similarity metric needs to measure concepts other than physical distance. In the context of similarity of texts, the correct approach is to measure the relative information gain between each other. Having read document A, how much more information can a reader get from reading document B?

A logical choice to measure this information gain is the Kullback-Leibler divergence (KL), which measures the difference between statistical distributions and is related to the Shannon and Wiener information theorems [KL51]. The more similar two documents and their probability distributions are, the less information is gained from one with respect to the other. Another option would be the Jensen-Shannon divergence [Lin91], which also measures the similarity between two probability distributions.

However, as the KL divergence is the metric used during training of the particular implementation [HBB10] used in this work, we keep it as measure of similarity between documents.

The KL divergence as a metric comes with two caveats:

First, the metric is not finite. The ratio of two probability distributions may incur a divide by zero issue. This can be remedied by adding a small amount ϵ to each component in order to prevent any division by zero. The value of ϵ then governs the upper numerical limit of the metric.

Second, the KL divergence is an asymmetric measure which is problematic when referring to true metric spaces as they assume the property of symmetry [Fré06]. However, the symmetry assumption is not universal in other domains, especially when looking at the application of human judgement to similarity [Tve77] and when a sense of hierarchy is subconsciously imposed by humans, such as for the example of saying "an ellipse is like a circle" rather than "a

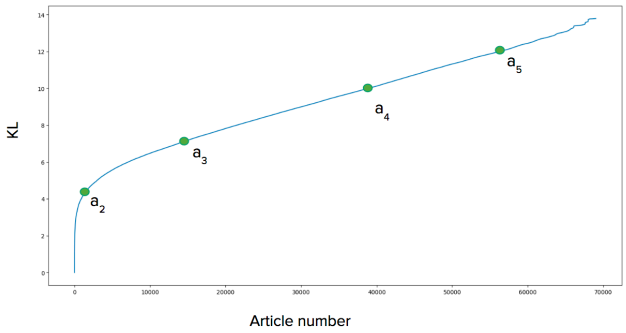

Figure 1: KL distribution of reference article a_1 against the rest of the articles in the corpus

circle is like an ellipse". The direction of asymmetry in our similarity space of news articles behaves in a similar way. If we have two articles talking about climate change for example, where one is a very detailed piece about climate change and the other is more of an overview, the information gained differs depending on the sequence that the articles are read in. Therefore we judge the KL divergence to deliver an adequate measurement of similarity between documents and, specifically, news articles.

To further evaluate the alignment of computed similarity with perceived similarity, we proceed with presenting a prototypical case of human-centric testing.

4 A Prototypical Case

In section 2 we discussed perception and the subjectivity of interpreting similarity by humans as well as how machines can compute similarity via different approaches with metrics like KL-divergence (section 3). In this section we describe a case following the method proposed in 2.3 to evaluate the alignment of similarity between humans and machines that helped us select the optimal model for the purpose of building content similarity recommenders for BBC News articles.

Once the articles have been translated into a distribution of topic probabilities, the KL divergence can then be used to rank articles by similarity. However, due to the fact that LDA is an unsupervised algorithm, it is difficult to measure the impact of adjusting the hyperparameters in contrast to supervised learning algorithms where loss and error provide a helpful constraint. Finding the optimal number of topics is particularly challenging when solely assessing the output topics and the similarity space the model spans.

Again, this is where the perceived similarity and human-centric tests show their strength. By comparing the similarity ranking of the model to the ranking performed by people through a variation on *triangle tests*, we provide a clear means to see which model conforms best to human judgement. This provides a way to deal with the key challenge in using LDA (or similar unsupervised learning methods): how to quantify the impact of tuning the hyperparameter reponsible for the number of topics.

4.1 Triangle Tests

We trained three LDA models with 30, 50 and 75 topics respectively, using 70 000 articles from BBC News Online published in 2017. From the set we selected a reference article a_1 and computed the KL divergence between the reference and all other articles in the set for one model. We then order the results from similar (small KL) to less similar in order to pick a diverse set of articles for testing. Figure 1 displays the distribution of articles ordered by KL between article a_1 and the rest of the articles in the corpus using the 30 topic model. Thus, we can select a set of articles (a_1 - a_5), to carry out the triangle tests.

The next step is to use the selected articles and create a questionnaire with sixteen questions. Each question contains three articles from the set: an anchor article and two comparative articles (A and B) that are located in different positions of the similarity space. The name for the test is drawn from the fact that three articles are always presented as mentioned in section 2.2. We asked ten journalists to read each anchor article alongside the two comparative articles. They then indicate which one, in their opinion, was more similar to the anchor article. The questions and order of the comparative articles were shuffled between participants.

The purpose of the test was to be able to compare the responses of the journalists with the responses of the different LDA models. Each model outputs a different KL value between articles depending on the hyperparameters (principally: number of topics) used. Therefore we expect different LDA models to have differing alignment with human judgement.

In order to evaluate the performance of the different models we calculated how many answers per participant agreed with the answers given by the model and therefore which model is best aligned with human interpretation. The results of this evaluation with 30, 50 and 70 topics models are displayed in Figure 2. When comparing the three models, the 50 topic model shows the best average alignment (70 percent) and least variance across the different testers. In general, all models show good alignment with human perception and certainly performs better than randomly selecting the correct answer, which is $\frac{1}{2}^{16}$. Additionally this also provides validation that human perception is highly aligned to our chosen similarity metric.

This gives confidence in the results obtained and

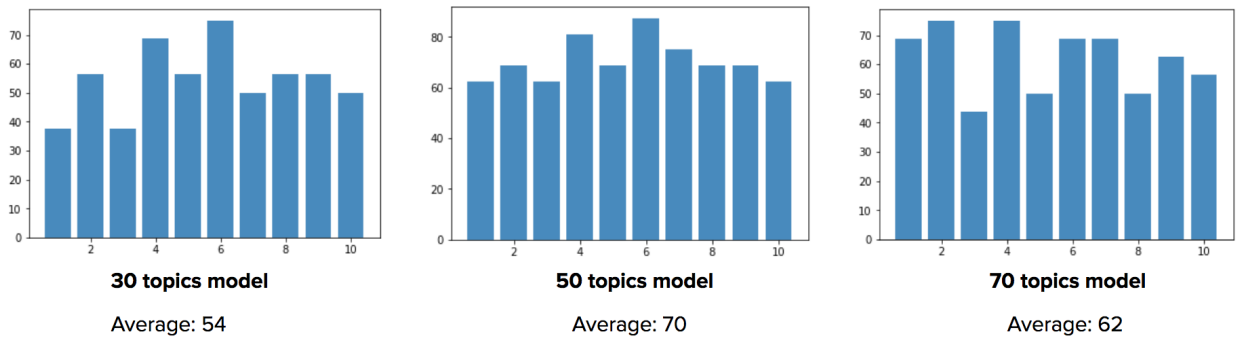

Figure 2: Percentage of answers aligned between the 30, 50 and 70 topic models and the respondents of the test. x-axis represents participant number, y-axis percentage of responses aligned with each model

allows us to proceed with the 50 topic model for a content similarity recommender in production.

5 Towards content similarity recommendations

With the best model selected, we can build an automatic *topic scoring* pipeline that, for every article published, transforms the article into a topic probability distribution. These distributions are persisted in a database and made available to the recommendation system. Using the KL divergence as the similarity metric, the recommendation system can calculate the similarity between each article pair and thus find the N most similar articles for a given article and serve them as recommendations. The recommended articles may be be further ranked and filtered according to business rules.

6 Conclusions and Future Work

The prototypical test shows the potential of this methodology in capturing alignment between human and machine perception of similarity. Additionally, it facilitates the selection of parameters for the LDA model. It has helped us discriminate between the three models and suggests the 50 topic model as the most appropriate. For pragmatism, we selected a limited number of articles and testers, however we believe these findings validate the use of this type of testing for general use and we consider this guidance for extracting stronger conclusions given a bigger sample.

In this contribution we have stated the need of measuring content-similarity in a news organisation with the motivation of building content similarity recommenders. We have revised methods to measure human and machine perception of similarity and presented a prototype of a human-centric test to evaluate the alignment between computed and human similarity with the purpose of assisting in the selection of parameters of the topic modelling algorithm LDA. The findings obtained show the strong potential of these types of tests. In the future we plan to apply the LDA model to build more sophisticated recommenders that takes into account the reading profile of users or sequential modelling.

References

[15301] ITU-R Recommendation BS. 1534-1. Method for the subjective assessment of intermediate quality level of coding systems, 2001.

[AM99] Cynthia M Aguilar and Douglas L Medin. Asymmetries of comparison. *Psychonomic Bulletin & Review*, 6(2):328–337, 1999.

[BGJM17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, (5):135–146, 2017.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[BRN99] John S Bradley, R Reich, and SG Norcross. A just noticeable difference in c50 for speech. *Applied Acoustics*, 58(2):99–108, 1999.

[CL95] Chun-Hsien Chou and Yun-Chin Li. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. *IEEE Transactions on circuits and systems for video technology*, 5(6):467–476, 1995.

[Fré06] M Maurice Fréchet. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 22(1):1–72, 1906.

[GBG+18] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan, May 2018. European Language Resource Association.

[GDF13] Florent Garcin, Christos Dimitrakakis, and Boi Faltings. Personalized news recommendation with context trees. In *Proceedings of the 7th ACM conference on Recommender systems*, page 105112. ACM, 2013.

[HBB10] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[HJ68] Errol R Hoffmann and Peter N Joubert. Just noticeable differences in some vehicle handling variables. *Human Factors*, 10(3):263–272, 1968.

[JGBM17] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

[KKGV18] Dhruv Khattar, Vaibhav Kumar, Manish Gupta, and Vasudeva Varma. Neural content-collaborative filtering for news recommendation. *NewsIR@ ECIR*, 2079:45–50, 2018.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[LDP10] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, pages 31–40, New York, NY, USA, 2010. ACM.

[Lin91] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[LM14] Q. Le and T. Mikolov. Distributed representations of phrases and their compositionality. In *International conference on machine learning*, pages 1188–1196, 2014.

[MCCD13] Tomas Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. volume Workshop Track, pages 1301–3781, 2013.

[OO85] MAPDE O'MAHONY and N Odbert. A comparison of sensory difference testing procedures: Sequential sensitivity analysis and aspects of taste adaptation. *Journal of Food Science*, 50(4):1055–1058, 1985.

[Pan95] Davis Pan. A tutorial on mpeg/audio compression. *IEEE multimedia*, 2(2):60–74, 1995.

[RP16] European Union Regulation and Protection. Regulation (eu) 2016/679 of the european parliament and of the council. *REGULATION (EU)*, 679, 2016.

[TASJ14] Michele Trevisiol, Luca Maria Aiello, Rossano Schifanella, and Alejandro Jaimes. Cold-start news recommendation with domain-dependent browse graph. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 81–88. ACM, 2014.

[Tve77] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[Wal92] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.

[YBDS+17] JM Yearsley, A Barque-Duran, E Scerrati, JA Hampton, and EM Pothos. The triangle inequality constraint in similarity judgments. *Progress in Biophysics and Molecular Biology*, 10, 2017.