Quote Clustering in Online News

Orr Keshet and Evan Rosen
{okeshet,erosen}@cs.stanford.edu

Abstract

The notion that information moves through social networks has been widely discussed[3], however, with the growing availability of large digital corpora, the ability to quantitatively model this phenomenon is new. To this end we explore a large corpus of online news quotations looking for cases of noisy reproduction and the factors which influence such noise. An essential step in this process is distinguishing mutational variants from entirely independent but similar quotations. Given that the question of whether two quotes really were derived from the same original utterance cannot be known with complete certainty, it is not immediately apparent how to make progress on such a task. Our appraoch is twofold: On the one hand we attempt to frame the problem in terms of supervised learning, annotating data using Mechanical Turk. On the other hand, we approach the problem more from the perspective of unsupervised clustering, projecting the data into a variety of metrics which allows us to test and extend our linguistic intuitions about the dataset.

1 Introduction

For the sake of this project, we restrict our attention to the inspection of quote distance metrics. Such distance metrics can serve as the basis for arbitrarily complex clustering algorithms, as in [5, 4], and they also provide a relatively direct way to test linguistic intuitions. In the spirit of the formal definition of a metric, we primarily test our distance metrics on pairs of quotes, ignoring the more global implications like transitivity for the moment.

The first component of this project involved sorting this large set of quote pairs by various distance metrics and hand-inspecting the results. In this way were were able to confirm and reject a several plausible hypothesis regarding mutations of online new quotes. In general, our results are mostly negative, suggesting that noisy reproductions in the news are less frequent than originally expected. While the rejection of such socialogical and linguistic assumptions is useful, we wanted to provide a systematic way to ensure that we were in fact looking at the right dimensions of variation, or to at least distinguish between the usefulness of the metrics which we proposed.

The design of a sensible evaluation scheme constitutes the second significant component of this work. This task presents two main obstacles. While a human reader consecutively reading two news articles on the same subject might easily tell when two different quotations are in fact variants of one another, it is much less obvious how to confidently reject two quotes as being variants of one another. The second issue arises more as a feature of the dataset. We only have access to the quotes themselves and thus cannot use any of the contextual information present in the article and webpage. Were we able to view the original new article in its entirety, it is likely that a human could make much more confident estimates about common quote origins.

Despite these callenges, we decided that human evaluation was still our best option and provides a reasonable baseline against which to compare our automated distance metrics. For example, a pair of quotes with edit distance one, which differ in only s single semantically distinct word, could easily be mistaken to be variants of the same quote by an edit distance metric, but would be immediately picked out by a human evaluator as semantically incompatible. We annotate quote pairs as plausible variants of a single original utterance using Amazon's Mechanical

Turk and then compare these annotations with the result of a rudimentary clustering algorithm based on each metric on its own.

2 Dataset

The dataset for this work is taken from a large corpus of mainstream news web sites and blogs, collected as part of the MemeTracker project [5]. This amounts to a large set of text files which, for each webpage record the url, date of publication, any quotations, and any outgoing url link, where the date of publication corresponds to the time at which the page was pushed to an RSS feed. While most of the pages are in English, a significant proportion are in Spanish and French and even several less common languages like Bahasa Indonesia.

Interestingly these rarer languages seemed to show up disproportionately in the set of close but non-identical quote pairs. For example, the following Indonesian phrases, which literally translates to *"a technique of buying a home without capital,"* all showed up many times in the corpus.

> *"teknik membeli rumah tanpa modal"*
>
> *"ada teknik membeli rumah tanpa modal"*
>
> *"mau tahu teknik membeli rumah tanpa modal"*
>
> *"mau tahuteknik membeli rumah tanpa modal"*

We suspect that this result might be due to a different use of quotation marks in other languages, as this quote, when taken literally, seems unlikely to have been a real persons statement. Another plausible explanation is that these quote vairations are the result of a less standardized orthography. This situation seems to arise in places in which the the national written language is learned as second language for many speakers. Another source of complication when dealing with foreign languages results from the fact that non-ascii characters have been replaced with spaces. All of this is to say that the problem of disambiguating quotes is very language specific and therefore we only focus on English quotes, though non-english quotes are still in the data, and make up edges in the link graph.

It is also worth noting that the dataset is extremely large. One month's worth of compressed quote data is approximately 1GB and consists of approximately 10 million webpages. We had access to data from 2008 to the present, but never constructed a single graph from the entirety of the quotes, due to time and memory constraints.

3 Methodology

3.1 Graph Generation

To better focus on requotation and reference in online news, we wanted to bias our dataset towards well connected web pages. That is, we looked for a collection of webpages which had at least one link to other webpages in the corpus. By filtering out any pages which did not contain any quotes or links to other pages in the corpus, we were able to significantly reduce the pages under consideration to approximately 300,000 per month. This significantly reduced the computational costs of our algorithm. Moreover, this filtering has the effect of reducing the number of connected components, making for more informative structural properties like shortest path length.

3.2 Test Pair Generation

The goal however is not just to generate a graph but to extract quote pairs. This requires first choosing a node pair and then a quote pair form the list of quotes associated with that page. As this space is combinatorially large, for most of our trials, we take random samples on the order of hundreds of thousands of unique pairs of form `<node:quote,node:quote>` . This approach is still somewhat problematic however, because we are often looking for quote pairs which are statistically unlikely. One solution, is to restrict the set of pairs to be bewteen pages published near one another. We stumbled upon this after noticing a worsening in results upon increasing the size of our test file. Because the test files are ordered by publication time, we were getting a small subset of articles all published within the same few hours of one another, which as we might expect, were about a much more homogenous set of topics.

While the generation of a million random quote pairs may have sufficed for our automated metrics, we obviously did not want to be giving mechanical workers entirely random pairs. Our initial approach was to use a simple automatic metric like character edit distance to sort the set of sampled pairs and then use only the most similar pairs for Mechanical Turk Evaluation. Unfortunately, this approach simply did not yield enough quote pairs which were remotely similar. After a significant amount of time spent tuning this sorting algorithm we came to the conclusion that the quote in which were interested were just less common than we had expected. By sorting the data by a variety of distance metrics and reruning the sample on different files of the course of several days in combination with grepping the data

files by hand, we finally assembled a meager set of 30 possible quote pairs which weren't so obviously unrelated that it would be foolish to as a turker to annotate them. This is the data set used for our evaluation discussion below.

3.3 Distance Metrics

We experimented with various metrics in order to determine if two quotes are in fact derived from the same original utterance. Our design choices in making these distance metrics were largely informed by linguistic ituition about misquotation. We incorporated these metrics into an rudimentary clustering technique that is based on thresholds. Additional deatils about this are forthcoming in section 3.5. Later, we evaluted these quote pair clustering against mechanical turk annotations. Throughout the paper, in the results section and the subsequent analysis, we dig deeper into each metric and analyse and compare one metric to others in the hopes that such an analysis will provide some linguistic insight.

3.4 Gold Standard Quote Pairs

We used Amazon's Mechanical Turk in order to generate a set of gold standard quote clusters. Mechanical Turk is frequently used by members of the NLP community for obtaining data that is both cheap and reliable[6]. We asked each worker to look at a set of quote pairs and determine whether a given pair of quotes, they are similar enough as to make them believe they originated from the same source. For each quote pair we had three unique workers submit answers and then used a merging algorithm to determine the concensus result. The merging algorithm simply looked for a majority vote. Workers were paid USD $0.05 per five quote pairs. We plot the distribution of consensus Mechanical Turk result in Figure 1.

Note that for the purposes of the evaluation we treated quote pairs annotated by workers as 'possibly' as 'no'.

3.5 Threshold Based Clusters

As noted above, a distance metric can be incorporated into an arbitrarily complex clustering algorithm. While we are not interested in the development of such algorithms here, we still need to use the gold standard quote pairs to evaluate our metrics. While we technically have a scale along which workers can place quotes ('yes', 'possibly', 'no') the consistency of this annotation is difficult to determine. To measure the similarity between turker annotation and automated metrics, we then used a simple binary classification scheme as the basis for evaluation.

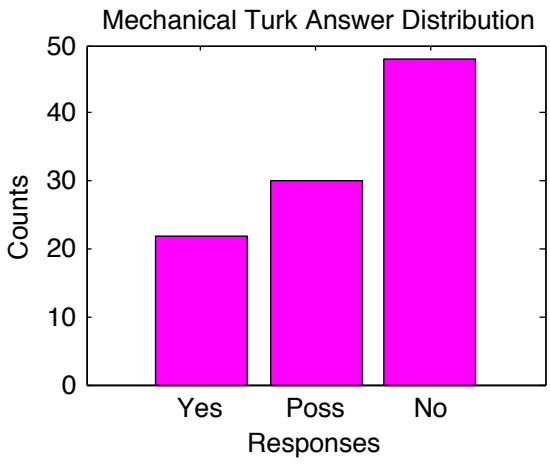


Figure 1: Mechanical Turk Results

To turn out distance metrics into classifications we just thresholded the metrics at the point which maximized the F1 score. While some sort of hold out set would have been ideal, the size of our data set made this infeasible. As an example, with a charcter edit distance threshold of 10 any two quotes less than 10 edits from one another are considered to be the same original quote. Using the pairs resulting from the gold standard annotation mentioned above, we can then compute precision and recall statistics for each metric. It is important to note that the pairings generated using this threshold method are in no way meant to be rigorous machine learning results but rather they are meant to serve as the basis for informally comparing our distance metrics.

3.6 Threshold Tuning

In the previous section we metioned that we created clusters based on quote pairs having some metric value below a certian threshold. In order to find this treshold value, in the way that maximizes the results against the annotated task, we sort the gold standard quote pairs and try out different threshold values on the bounadires between the different pairs. At the end, we are left with the treshold value that partitions the cluster in a way that is maxmizes the F1 score.

4 Results

We present the precision, recall, and F1 scores in Table 1.

5 Discussion

In this section we discuss and evaluate the results presented in the previous section. The discussion

Metric	Prec	Recall	F1
CharEdit	0.50	0.33	0.4
WordEdit	0.71	0.45	0.56
ShortestPathLen	0.63	0.5	0.56
Temporal Diff	0.18	0.25	0.21

Table 1: Metric Performance

starts off with analysis that is specific to each metric and then follows with some general notes on the types of quotation variants which these metrics identified. As we discuss each metric, our goal is to dig deeper and reveal some linguistic insights and see how these in turn help better explain the results.

5.1 Metric Specific Analysis

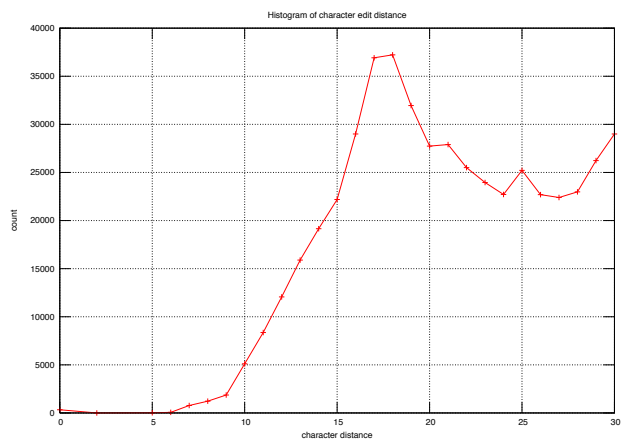
5.1.1 Character and Unigram Edit Distance

As one can observe in Table 1, the character edit distance metric performs worse than others. To investigate possible reasons as to why that is the case we must begin with the ituition that motivated us to use this metric in the first place. With a metaphor of genetic mutation in mind, we expected many of the variants to be due to unintentional reproduction errors such as typos or small grammatical mistakes. Such differences can be captured well by character edit distance. Consider the table and plot shown in Figure 2. Note the sparsity in counts for character edit distances less than 7. When generating this plot, we recorded edit distances for over 300,000 quote pairs so it is quite astonishing that there are barely any misquotations in the form of spelling errors or other reproduction errors.

If we conceptualize quote reproduction as a game of telephone, this result is sensible. For, if we mishear someone, we rarely invent entirely new words to suit the acoustic signal, but rather map this acoustic signal to the some set of preexisting words, with which we are familiar with. Realistically this is probably the result of automatic spell checkers which have now become ubiquitous. Counts begin to pick up around edit distance 7, however suggesting that misquotions occur at the word level (which makes sense given the average word length in our corpus was 4.5). Whether this process is occuring at the point of human interpretation or through automatic spell checkers, this is an essential lesson to take from exploration of quote variants: changes occur in some window of granularity. To small a granularity and the new variant will be quickly recognized as non-sense, too large a granularity and the very topic of a quote may be lost. As a follow up, we decided to investigate this suggestion by implenting a word level edit distance metrics. The

Count	Character Edit Distance
0	336
1	0
2	3
3	0
4	0
5	19
6	66
7	733
39	65653

(a) Character Edit Distance Counts

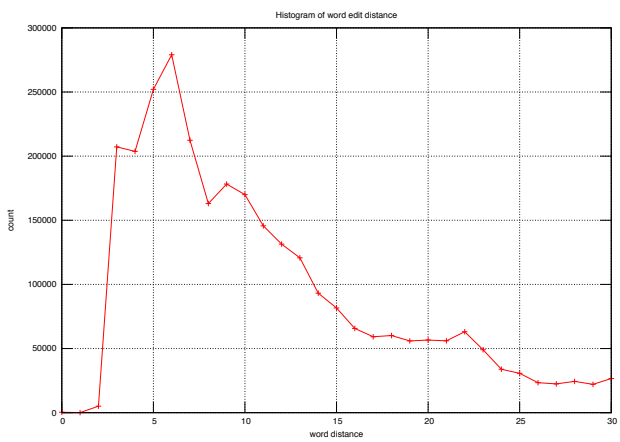


(b) Character Edit Distance Histogram

Figure 2: Character Edit Distance Analysis

Count	Word Ed	Norm. Char Ed
0	336	336
1	45	24
2	5050	9036
3	207228	77656
4	203706	162826
5	252123	125287
25	30652	—s

(a) Word Edit Distance Counts



(b) Word Edit Distance Histogram

Figure 3: Word Edit Distance Analysis

histogram for unigram word edit distances is shown in Figure 3.

As one can observe, there is a much greater frequency of quote pairs with lower word edit distance than character edit distance, but this is simply the result of a single word edit corresponding to multiple character edits. In an attempt to compare these distributions more meaningfully, next to each word edit distance d_w, we aggreggated the number of quote pairs with the character distance d_c such that $5(d_w - 1) < d_c < 5d_w$. For the first few word edit entries, this serves as an estimated lower bound on the number of necessary word edits. If a pair of quotes are 3 character edits apart, then they can be no less than 1 word edit distance, and two quotes are 7 character edits apart, then they are unlikely to be less than 2 word edits apart.

A key question to answer with this comparison is whether most of the variation is in affixes like conjugation or case marking, or whether these changes are full words. The fact that the aggregated character edit counts tend to be higher is likely a result of character edit distance finding a quicker path of edits between the two quotes than word edit distance due to overlapping substrings. Take as an example, the

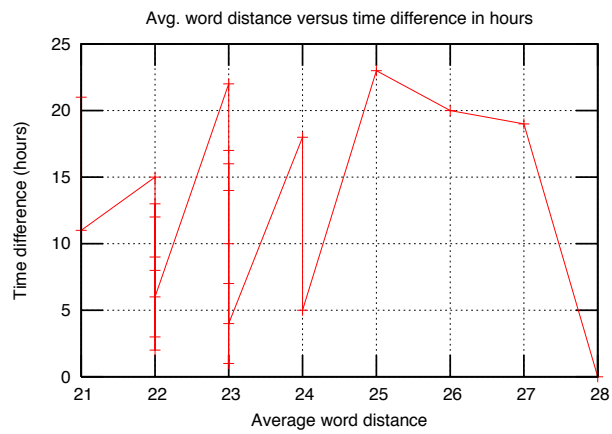


Figure 4: Temporal Distance vs Average Word Distance for Large Size Data

quote pair,

> *"What's the point if you can't extort"*
>
> *"What they like most about the holidays is"*

.While very different in words we can easily transform *"the"* in to *"they"* without making 4 character edits. In cases like this however, the character edit distance strays significantly from the semantic similarity of the quotes. When a human reads those two quotes, the similarity between *"the"* and *"they"* does not make the quotes any more likely to variants of one another. For this reason, we would expect word edit distance to do better when compared to Mechanical Turk evaluation and we can see that this is case by looking at Table 1.

5.1.2 Temporal Distance

The last metrics we use attempted to analyze the signficance of temporal relationship between article publication. We initially assumed that the shorter the time between one publicatio of a quote and another, the higher the fidelity of the reproduction. The results for this metric are the worst of any metric we experimented with when evaluated against Mechanical Turk annotations, though this is not surprising given the coarseness of the distance metric. Nonetheless, we were able to draw some interesting conclusions from the general relationship between temporal distance and linguistic metrics.

Figure 4 shows the correlation between temporal distance and word edit distance for a very large dataset containing about a million quote pairs. This figure suggests that the sheer amount of quotes generated is so large and so diverse that considering the timestamp on the quote alone is not a good predictor for similar quotes. This explains the poor results

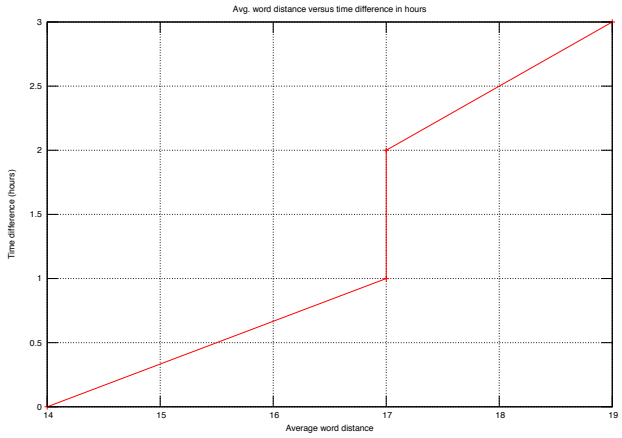
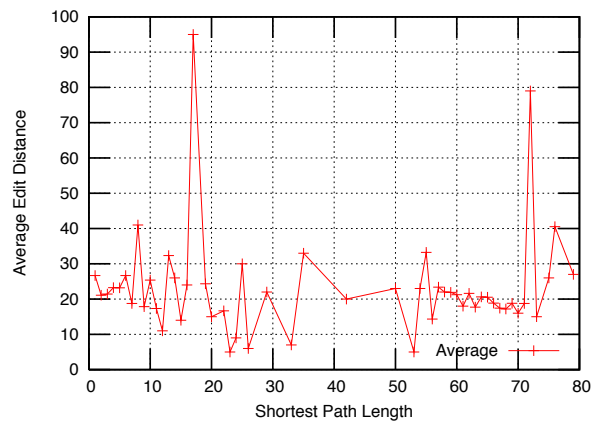


Figure 5: Temporal Distance vs Average Word Distance for Specific Event Quotes

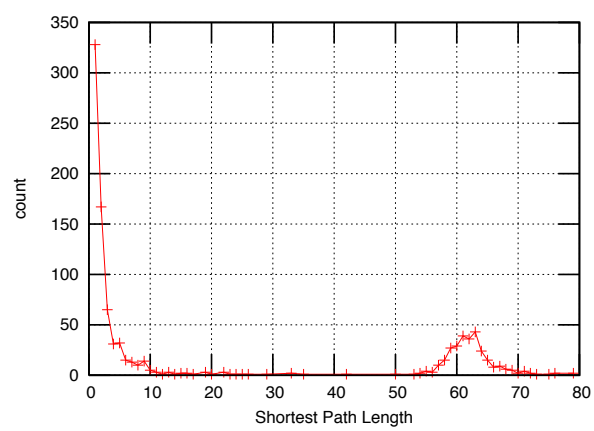
previously shown. Even so, one would still expect that if two quotes were somewhat similar in content (i.e. the pair was not randomly chosen), the fact that they were reproduced in close proxmity might indicate that they are infact reproduction of the same utterance. We decided to manually gather quotes from a specific news event and plot the temporal difference as a function of average word edit distance to see if we can force the quotes to be similar enough as to see how well does temporal distance predict word distance. This figure in fact confirms our expectation that temporal difference should be useful when combined with others.

5.1.3 Shortest Path Length

On the intuition that pages near one another on the link graph might be more like to discuss the same topic, we examined the effect of link distance on quote similarity. We brought two basic assumptions to this question. On the one hand, we expect that pages which are close in the link graph are more likely to be topically related, say discussing a similar topic, and therefore more likely to contain related quotes. However because we are looking for non-identical pairs, we also expect something like norms to have an effect on local inconsistencies. In other words, if two pages are near one another on the link graph, then their authors are more likely to read the other's page, and mimic the variant which the other used. Figure 6a shows the average edit distance per link graph shortest path length. However, the odd shape in the middle of the graph suggested that many of the path length bins were sparse, which we confirmed upon closer examination. We plot the histogram of link graph shortest path lengths in Figure 6b.



(a) Path Length vs Word Edit Dist



(b) Path Length Distribution

Figure 6: Path Length Figures

5.2 General Analysis

While our aim in using mechanical turk was to introduce an objective baseline againts which to compare our metrics, in the end it may have introduced more noise than it was worth. It is important to note that the results reported in Table 1 are very much dependent on the worker's responses. It is not entirely clear whether the mechanical turk results should be treated as proper gold standard annotation for sever reasons. First, we should note that on many of the quote pairs, the workers disagreed with our proposed answer. This means that the workers answered in a way that we did not predict. Second, despite our attempt to catch workers who abused the system by having multiple workers annotate each quote pair, it seems that many workers selected an answer far too fast than possible if they really gave it just consideration. Many workers took less than 10 seconds to annotate 5 independent quote pairs.

Ultimately the product of this work has been to familiarize ourself with the very question of quote disambiguation. The main result from this has been to recognize that these variations are in fact far less frequent than originally expected, and partly due to the sparisty, especially difficult to resolve as variants of one another. Perhaps the most interesting finding have been anecdotal. For example, we found several quotes which differed only in stutterings or hesitation sounds. For example consider the pair:

> *"there was no rage or anything he was just like a robot stabbing"*
>
> *"there was no rage or or anything he was just like a robot stabbing the guy"*
>
> or
>
> *"vamos a eliminar todas las escuelas primarias precarias"*
>
> *"vamos a eliminar todas las escuelas primarias eh precarias"*

The quotes are almost surely in reference to the same original event, but likely correspond to different transcriptions. Somewhat reassuringly, these are exactly the type of changes which we admit journalists the right to make.

An even more interesting case of quote similarity is a pair which involves a reasonably significant grammatical transformation

> *"when he was attacking him he was calm it was like he was at the beach"*
>
> *"when he was attacking him he was calm like he was at the beach"*

After a bit of internet video sleuthing, it became clear that this example was actually from two different interviews of the same person, regarding a brutal decapitation on a trans-Candadian Greyhound bus [1, 2]. This possibility is especially interesting case, because it brings up the question of where the mutation occured. It might have been in the speakers head, in the transcription of an interview, or in the quoting of a news article by a blog. These different contexts for mutation offer an interesting view into the substructure of this problem.

6 Future Work

This work is primarily exploratory in nature; providing experience with the data such that we can create more realistic hypotheses. Perhaps the most significant task ahead of us is to reframe the scale on which we search for quote pairs. With more time, we would have liked to create link graphs spanning multiple months and sample even more pairs of quotes. The very question of the relationship between the frequency of non-identical quote variants and the number of pairs examined will be an interesting question to further investigate. In the case where a more exhaustive search does in fact result in more quotes, we will also look into a more rigorous use of gold-standard annotation. As we have framed the problem now, it is essentially one of binary classification for which we could use any number of more complex algorithms. This step is especially important, because it will allow use to easily combine information from the webgraph and the linguistic data into a single estimate of quote identity.

References

[1] http://news.bbc.co.uk/2/hi/americas/7535840.stm?ls.

[2] http://www.cbc.ca/canada/story/2008/07/31/greyhound-transcanada.html.

[3] Richard Dawkins. *The selfish gene*. Oxford University Press, New York, 1976.

[4] Andreas Krause Jure Leskovec, Manuel Gomez-Rodriguez. Inferring networks of diffusion and influence. *KDD*, 2010.

[5] Jure Leskovec, Lars Backstrom, and Jon M. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.

[6] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*, pages 254–263, 2008.