Detecting Rumor and Disinformation by Web Mining

Boris Galitsky,
Knowledge-Trail Inc. San Jose CA 95127 USA
bgalitsky@hotmail.com

Abstract

A method to detect if a given text is a rumor or disinformation is proposed, based on web mining and linguistic technology comparing two paragraphs of text. We hypothesize about a family of content generation algorithms which are capable of producing disinformation from a portion of genuine text. We then propose a disinformation detection algorithm which finds a candidate source of text on the web and compares it with the given text, applying parse thicket technology. Parse thicket is graph combined from a sequence of parse trees augmented with inter-sentence relations for anaphora and rhetoric structures. We evaluate our algorithm in the domain of customer reviews, considering a product review as an instance of possible disinformation. It is confirmed as a plausible way to detect rumor and disinformation in a web document.

Introduction

Information that is published on the web and propagates through social networks can carry a lot of false claims. Published once, it can be copied into multiple locations with some edits and make an impression that multiple sources confirm untrue facts and fake opinions. Such fake information, rumor or disinformation may be distributed to manipulate public opinion; therefore its sources and posting of its various versions needs to be identified as fast as possible.

A fast growth of online information sharing media has made it possible for rumor to spread rather quickly. Unreliable sources can quickly spread inaccurate and intentionally false information in large quantities, so it is crucial to design systems to detect both misinformation and disinformation at the time it is indexed by search engines, included in feeds, etc.

In this study we are concerned with high volume of disinformation, assuming it is created and distributed automatically. It is hard to scale manual writing process and manual distribution, so for real attempts to manipulate public opinion we expect automated agents to create content (Galitsky & Kuznetsov 2013). To do that at a scale, they would have to obtain publicly available content and substitute some entities and their attributes in some manner. As a result, high quantities of a strongly opinionated content can be systematically created and distributed in favor of a certain group of people. The working assumption is that a certain content source would be exploited by such agents, given their mission. These agents take genuine content, substitute certain entities in favor of their mission, and distribute it. Moreover, the agents are expected to do some text re-phrasing to avoid easy detection of the real sources.

The key in handling these cases of disinformation would be to identify the source and highlight the substituted entities. Currently available copyright detection software is not well suited to do this job because of the potential high number of substituted entities. Hence the similarity between the fake content and original content is expected to be too low for copyright algorithms to determine.

The idea of publishing similar portions of information in various places to affect the public opinion is nicely expressed in the following quote:

"See, in my line of work you got to keep repeating things over and over and over again for the truth to sink in, to kind of catapult the propaganda." George W. Bush - 43rd US President

One can see how this procedure can be automated by taking a piece of information, rewriting it multiple times (which is entity/attribute substitution in our case) and publishing it in multiple places

“Political language. . . is designed to make lies sound truthful and murder respectable, and to give an appearance of solidity to pure wind.” George Orwell.

Instead of relying on social network topology information to track the sources and propagation of disinformation and rumor, in this work we rely on linguistic means to perform a similarity assessment between a given text and a candidate for its source on the web. The finding procedure of textual sources is conducted via web mining, employing search engine APIs.

According to (Mintz 2013), the best ways to find if information is factual is to use common sense. A reader should verify if a piece of information makes sense, if the founders or reporters of the sites are biased or have an agenda, and look at where the sites may be found. It is highly recommended to look at other sites for that information as it might be published and heavily researched, providing more concrete details. The art of producing disinformation is based on the readers' balance of what is truth and what is wrong. Hence most of the entities and their attributes, appealing to the former, are retained, and those referring to the latter are substituted.

Readers must have a balance of what is truth and what is wrong. There is always a chance that even readers who have this balance will believe an error or they will disregard the truth as wrong. (Libicki 2007) says that prior beliefs or opinions affect how readers interpret information as well. When readers believe something to be true before researching it, they are more likely to believe something that supports their prior thoughts. This may lead readers to believe disinformation.

Examples of disinformation as entity substitution

We use an example of well-known disinformation to analyze how it can be potentially scaled up.

In early 2007 Wikipedia community was appalled when an active contributor (believed" by the site to be a professor of religion with advanced degrees in theology and canon law), was exposed as being nothing more than a community college drop-out. The person at the center of this controversy was "Essjay" (Ryan Jordan), a 24-year-old from Kentucky with no advanced degrees, who used texts such as Catholicism for Dummies to help him correct articles on the penitential rite and transubstantiation.
http://educate-yourself.org/cn/wikipedialies20jan08.shtml

What we observe here is that substituting certain entities in popular religious texts, one can produce scholarly articles.

"On 25 September 2014, REN TV's website ran a story quoting Russian-backed insurgents as saying that "dozens" of bodies had been discovered in three graves, some with organs removed. It illustrated the story with an image of men carrying what appears to be a body bag.

Four days later, REN TV's website - from which a screen grab was taken, above - reported that "bodies continue to be discovered" in areas that it said had been recently vacated by Ukraine's National Guard. The report contained an image of numerous body bags placed on the ground near to what appears to be a piece of white wreckage.

But both of these images were details from photographs that had appeared over a month earlier on the website of the airline crash investigators:"
http://www.bbc.co.uk/monitoring/russian-tv-uses-crash-pictures-in-mass-grave-report

What has been done here is substitution of the attribute location and reason. The main entity multiple bodies have been reused, together with the associated image. What the agent, REN TV channel, did is substituted the values of location = 'airliner crash site' with 'area vacated by Ukraine's National Guard' and reason = 'airliner crash' with 'activity of Ukraine's National Guard'. The purpose of this disinformation is to produce negative sentiment about the latter. In this particular cases the fact of disinformation has been determined by the reused authentic image; however considerations of this paper is reused text. To perform the detection, we take a text (image, video or other media) and try to find a piece of similar content available on the web at an earlier date.

A high level view of a hypothetical disinformation creation tool

To be able to identify text containing rumor and disinformation, we need to hypothesize about a tool which would create it in arbitrary domain

For an efficient rumor producing tool, it needs some relevance machinery to filter content suitable to be included in the resultant text on one hand, and also a mechanism to track the rhetoric structure of the produced text, for example, by copying it from the source. One needs to use a high level discourse structure of human-authored text to automatically build a domain-dependent template for given topic, such as event description, biography, political news, chat and blog. In case of a dialogue or a text containing some kind of argumentative structure, this template is based on a sequence of communicative actions. In a general case we follow a certain epistemic structure extracted from multiple texts in a particular domain (for example, for a music event we present a performer biography, previous concerts, previous partnerships, and future plans).

A typical creative writing activity of an average author is searching and browsing the web for relevant information, then finding pieces and merging them together, followed by final text polishing. The objective of the rumor creation tool would be to simulate human intellectual activity while writing an essay, searching the web for relevant content and combining it in a proper way. We would expect the rumor creation tool to focus on final acceptance /rejections of candidate text fragments and making sure the overall writing is cohesive. The substitution mapping needs to be set up manually, such as actor1 → aggressor {list-of-names}, actor2 → victim { list-of-names }, attributes →means {weapon}.

Today, original content is created by human writers and therefore costly, slowly produced. Finding a way to automate content creation so that the result is satisfactory for human content consumers and perceived as original by

search engine isa key for a rumor creation tool. For web-based content generation, the relevance of formed sentences to the seed sentence is essential. A number of attempts to reformulate a text for the purpose of making it original are well known to search engines on one hand and produce uninteresting and semantically non-cohesive content even at the single sentence level.

The idea of web mining is that everything on the Earth has already been formulated and written in terms of style, and the task is to identify all facts, phrasing and opinions consistent with each other and combine them in a plausible flow. Our assumption for content generation is that it is impossible in most cases to really invent new phrase: something similar linguistically (but with different entities) has been posted somewhere on the web, so the task is two-fold:

1) find it

2) substitute entities from seed sentences in the mined sentences and merge them.

It is assumed in the body of text generation literature that learning a single topic-specific extractor can be easily achieved in a standard classification framework. However, we evaluate texts generated with and without parse tree learning and observed rather poor relevance of sentences mined on the web. For example, mining for the biography of Albert Einstein, we get 'Albert Einstein College of Medicine' result which should be filtered out, whereas 'Albert Einstein talent to foresee' search result is a legitimate expression to be included in a generated biography text.

Disinformation generation algorithm

We start with the seed (Fig. 1), one or multiple sentences each of which will form one or more paragraphs about the respective topics. These seed sentences can be viewed as either headers or informational centroids of content to be generated. We now iterate through each original sentence, build block of content for each and then merge all blocks, preceded by their seed sentences, together.

To find relevant sentences on the web for a seed sentence, we form query as extracted significant noun phrases from this seed sentence: either longer one (three or more keywords, which means two or more modifiers for a noun, or an entity, such as a proper noun). If such queries do not deliver significant number of relevant sentences formed from search results, we use the whole sentence as a search engine query, filtering our content which is duplicate to the seed.

The formed queries are run via search engine API or scraped, using Bing, Yahoo API or Google, as well as their /news subdomains depending on the topic of generated content; search results are collected. We then loop through the parts of the snippets to see which sentences are relevant to the seed one and which are not. If only a fragment of sentence occurs in the snippet, we need to go to the original page, download it, find this sentence and extract it.

For all sentences obtained from snippets, we verify appropriateness to form a content on one hand, and relevance to the seed sentence on the other hand. Appropriateness is determined based on grammar rules: to enter a paragraph cohesively, a sentence needs to include a verb phrase and/or be opinionated; mental space of cohesive information flow has been explored, for example, in (Galitsky et al 2008). Relevance is determined based on the operation of syntactic generalization (Galitsky et al 2010), where the bag-of-words approach is extended towards extracting commonalities between the syntactic parse trees of seed sentence and the one mined on the web. Syntactic generalization allows a domain-independent semantic measure of topical similarity between a pair of sentences, without it combination of sentences mined on the web would not form a meaningful text.

In addition to syntactic generalization, the tool verifies common entities between seed and mined sentence, and applies general appropriateness metric. The overall score includes syntactic generalization score (the cardinality of maximal common system of syntactic sub-trees) and appropriateness score to filter out less suitable sentences. Finally, mined sentences are re-styles and re-formatted to better fit together, and joined in paragraphs.

The content generation flow for the hypothetical content generation algorithm is as follows:

For sentence "*Give me a break, there is no reason why you can't retire in ten years if you had been a rational investor and not a crazy trader"*

We form the query for search engine API: *+rational +investor +crazy +trader*

From search results we remove duplicates, including *"Derivatives: Implications for Investors | The Rational Walk"*.

From the search results we show syntactic generalization (Galitsky et al 2012) results for two sentences:

```
Syntactic similarity: np [ [IN-in DT-a JJ-* ],  [DT-a JJ-* JJ-crazy ],  [JJ-rational NN-* ],  [DT-a JJ-crazy ]] 0.9
```

Rejected candidate sentence: Rational opportunities in a crazy silly world.

```
Syntactic generalization: np [ [VBN-* DT-a JJ-* JJ-rational NN-investor ], [DT-a JJ-* JJ-rational NN-investor ]] vp [ [DT-a ],  [VBN-* DT-a JJ-* JJ-rational NN-investor ]] 2.0
```

Accepted sentence: I have little pretensions about being a so-called "rational investor".

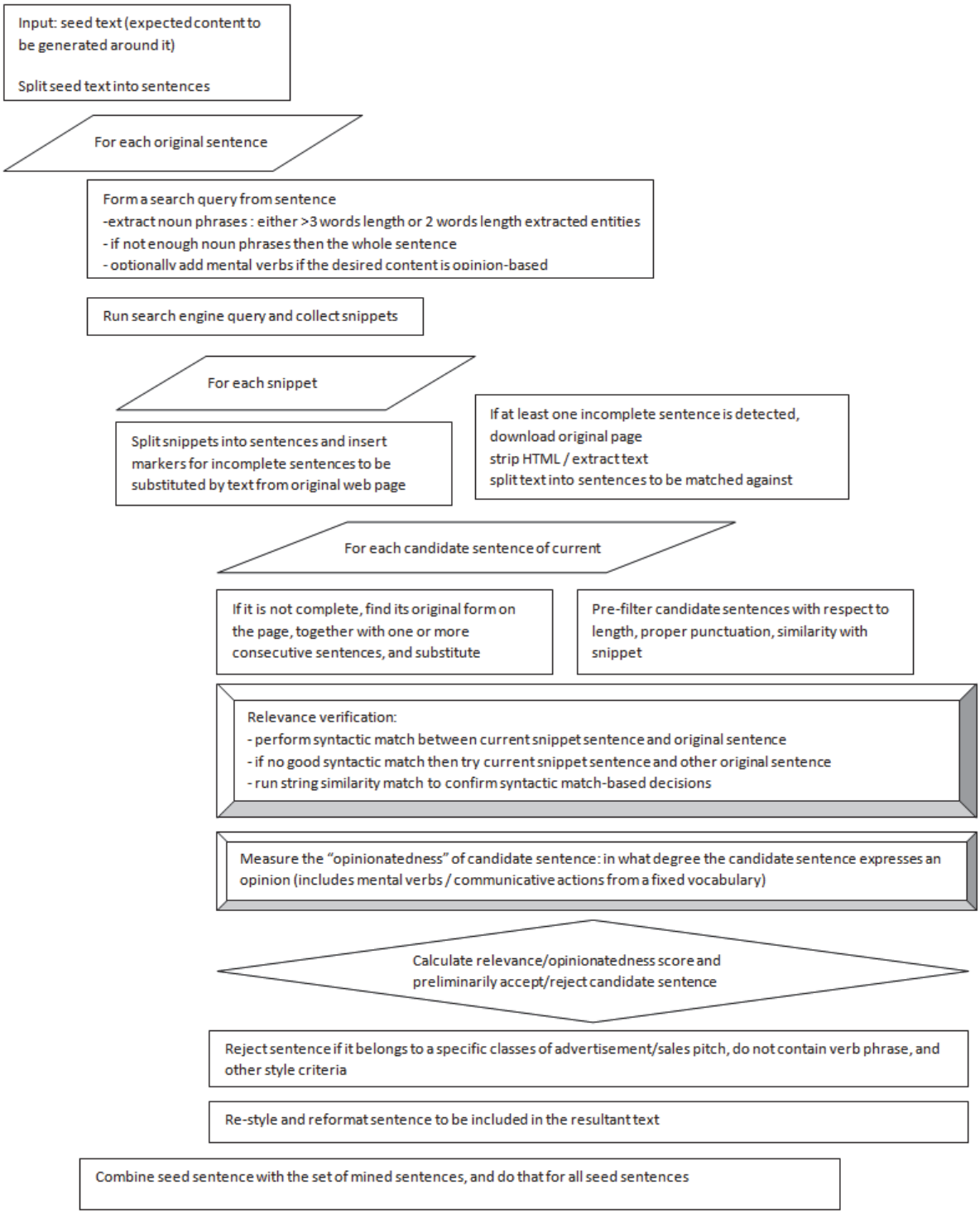


Fig. 1: The chart for a family of web mining-based content generation algorithms

As the reader can see, the latter sentence has significantly stronger semantic commonality with the seed one, compared to the former one, so it is expected to serve as a relevant part of generated content about "*rational investor"* from the seed sentence.

Disinformation and rumor detection algorithm

Input : a portion of text (possibly published on the web)

Output: categorization of input text as normal or disinformation (also including the original authentic information, and its source)

1. For a given portion of text (seed), find most significant sentences (similar to summarization)
2. For each of the most significant sentences, form an query in the conjunctive form.

$$(X_1 \vee Y_1) \wedge (X_2 \vee Y_2) \wedge \ldots \wedge (X_n \vee Y_n)$$

Where *Xi* and *Yi* are keywords, some of them are expected to be substituted so they will not occur in a potential search result

3. Run the search and collect all search results for all queries.
4. Identify common search results for the set of queries
5. Form the set of candidate texts which could be a source for the texts being analyzed
6. For each candidate, compare it with the seed. If high similarity is found, along with the substituted entity, then disinformation is found.
7. Identify the mapping of entities and their attributes from the seed text to the source text. Highlight substituted entities and attributes
8. Identify sentiments added to the seed text compared to the source.

Steps 1 to 5 are straight-forward, and 6-8) require a linguistic technology to match two portions of text and map entities and their attributes.

Linguistic technology which recognizes disinformation content needs to be developed hand-in -hand with content generation linguistics. If a content generation algorithm does rephrasing on the sentence level, applying parse tree-based representation, then a recognition algorithm needs at least as detailed linguistic representation as parse trees. Furthermore, if a content generation algorithm relies on inter-sentence level discourse structure, it needs to be represented by a detection algorithm as well.

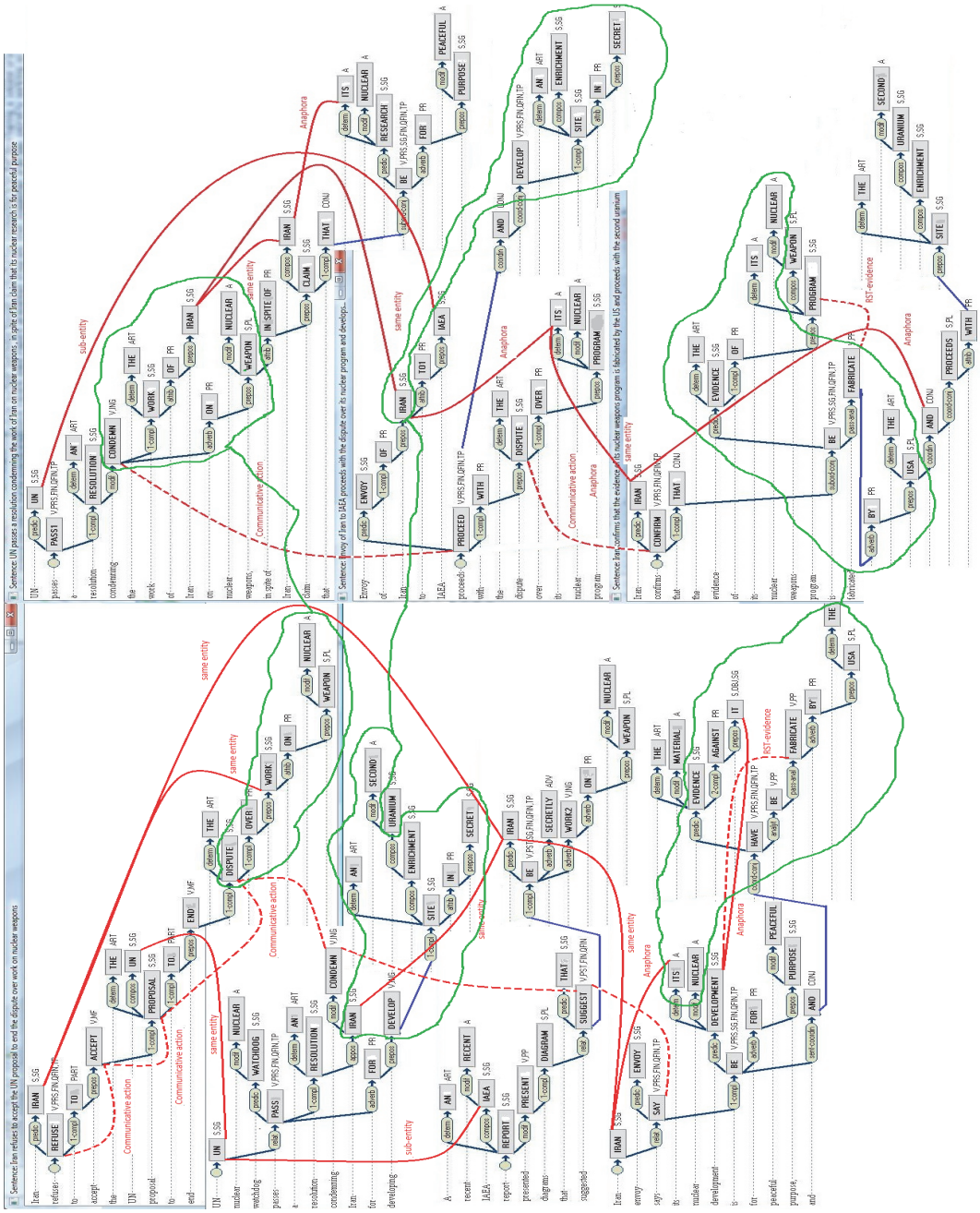
The results of the content generation family of technologies presented in this paper are not detected by search engines at the time of writing. This is due to the belief that they do not use parse tree – level representation for sentences in a search index. Once search engines employ parse tree representations, content generation algorithms would need to be capable of modifying rhetoric structure of text at the paragraph level to avoid being detected.

Matching seed and source texts

For two portions of text, we want to establish mapping between corresponding entities and their attributes. To do that, we need to employ parse trees as well as discourse relations, to form a parse thicket for a paragraph (Galitsky et al 2012, Galitsky 2013; 2014). Formally, the matching problem is defined as a generalization operation, finding the maximum common subgraph of the parse thickets as graphs. In this paper we provide an example of matching the seed,

"Iran refuses to accept the UN proposal to end the dispute over work on nuclear weapons",
"UN nuclear watchdog passes a resolution condemning Iran for developing a second uranium enrichment site in secret",
"A recent IAEA report presented diagrams that suggested Iran was secretly working on nuclear weapons",
"Iran envoy says its nuclear development is for peaceful purpose, and the material evidence against it has been fabricated by the US",
against candidate source
"UN passes a resolution condemning the work of Iran on nuclear weapons, in spite of Iran claims that its nuclear research is for peaceful purpose",
"Envoy of Iran to IAEA proceeds with the dispute over its nuclear program and develops an enrichment site in secret",
"Iran confirms that the evidence of its nuclear weapons program is fabricated by the US and proceeds with the second uranium enrichment site"
The matching results are as follows:

[NN-Iran VBG-developing DT- NN-enrichment NN-site IN-in NN-secret]*
*[NN-generalization-<UN/nuclear watchdog> * VB-pass NN-resolution VBG condemning NN- Iran]*
[NN-generalization-<Iran/envoy of Iran> Communicative_action DT-the NN-dispute IN-over JJ-nuclear NNS-
[Communicative_action - NN-work IN-of NN-Iran IN-on JJ-nuclear NNS-weapons]
[NN-generalization <Iran/envoy to UN> Communicative_action NN-Iran NN-nuclear NN- VBZ-is IN-for JJ-peaceful NN-purpose],*
*Communicative_action - NN-generalize <work/develop> IN-of NN-Iran IN-on JJ-nuclear NNS-weapons]**



Preliminary evaluation

We collected a set of thousand product recommendations and consider them as "disinformation" relative to the product features descriptions by the manufacturers and retailers. Given a set of product queries, we obtained a few opinionated texts on each.

https://code.google.com/p/relevance-based-on-parse-trees/downloads/detail?name=Queries900set.xls.

This opinionated text, such as an amazon review for a digital camera, we then submitted as a query against formal product descriptions. The other sites we mined for imitations of "rumor" content are review sites, Yahoo answers, and topic-specific sites containing reviews. For the source content, we use all sites on the web

In the context of our evaluation, the opinionated data can be viewed as potentially being a rumor, and actual product description is a source. The attribute substitution occurs by altering some parameters of the product:

- the consumer who wrote a review has a different estimate of a parameter from a manufacturer;
- the consumer specified a product attribute/feature which is lacking in product description
- the consumer adds sentiments related to product attributes and usability.

The task is to identify the proper source (product description) on the web along with the set of substituted attributes. Hence we believe our evaluation domain is relevant to an actual disinformation domain in terms of web mining properties and its linguistic features.

Table 1: evaluation of finding source text on the web

Seed Text fragments /size	Recall of finding source page, %	Precision of finding source page,%	Substituted attributes found, %	Sentiments found, %
Single sentence, <15 words	71.2	67.2	78.9	62
Long compound sentence, >15 words	67.4	73.3	71.6	70.1
2-3 sentences	72.9	72.1	65	64.5
4-5 sentences	70.4	80.6	62.7	61.3

We automatically formed the Seed Text dataset by mining the web for opinions/reviews. It includes 140 seed texts, from simple sentences of less than fifteen words to a fairly detailed multi-sentence product review. The size of the seed needs to correspond to the size of the identified source portion of text

We manually reviewed the rumor finding sessions and made assessments of precision and recall (Table 1). Once can see that the more information we have in the seed (the longer the text), the higher the precision of rumor identification procedure is, and the lower the percentage of identified attributes is. Recall and the proportion of identified sentiments do not significantly depend on the size of seed text.

Conclusions and Related Work

We were unable to find a systematic source of disinformation on the web. However, opinionated data on user products being related to product descriptions, turned out to be an adequate way to evaluation of out algorithm. We confirmed that it performs fairly well in identifying textual sources on the web, entity substitution and sentiment detection. Our evaluation addressed the cases of various complexities of text and demonstrated that disinformation can be detected varying from a single sentence to a paragraph containing up to five sentences (having entities substitution distributed through this portion of text).

(Seo et al 2012) focused on two problems related to mitigation of false claims in social networks, based on the source topology rather than linguistic approach. First, the authors study the question of identifying sources of rumors in the absence of complete provenance information about rumor propagation. Secondly, they study how rumors (false claims) and non-rumors (true information) can be differentiated. The method is based on an assumption that rumors are initiated from only a small number of sources, whereas truthful information can be observed and originated by a large number of unrelated individuals concurrently. Unlike the current approach based on web mining and linguistic technology, the authors rely on utilizing network monitors; individuals who agree to let us know whether or not they heard a particular piece of information (from their social neighborhood), although do not agree to let us know who told them this information or when they learned it.

Besides social network analysis, cognitive psychology helps identify the cognitive process involved in the decision to spread information (Kumar & Geethakumari 2014). This process involves answering four main questions viz consistency of message, coherency of message, credibility of source and general acceptability of message. We have used the cues of deception to analyse these questions to obtain solutions for preventing the spread of disinformation.

(Canini et al 2011) studies indicated that both the topical content of information sources and social network structure

affect source credibility. Based on these results, they designed a novel method of automatically identifying and ranking social network users according to their relevance and expertise for a given topic. Empirical studies were performed to compare a variety of alternative ranking algorithms and a proprietary service provided by a commercial website specifically designed for the same purpose.

(Qazvinian et al 2011) address the problem of rumor detection in microblogs and explore the effectiveness of 3 categories of features: content-based, network-based, and microblog-specific memes for correctly identifying rumors. The authors showed how these features are effective in identifying the sources of disinformation, using 10,000 manually annotated tweets collected from Twitter. In the current study, a deeper linguistic means are required to identify larger portion of text with disinformation.

A broad range of methods has been to study the spread of memes and false information on the web. (Leskovec et al. 2009) use the evolution of quotes reproduced online to identify memes and track their spread overtime. (Ratkiewicz et al., 2010) created the "Truthy" system, identifying misleading political memes on Twitter using tweet features, including hashtags, links, and mentions. Other projects focus on highlighting disputed claims on the Internet using pattern matching techniques (Ennals et al., 2010).

Instead of identifying rumors from a corpus of relevant phrases and attempting to discriminate between phrases that confirm, refute, question, and simply talk about rumors of interest, we apply a paragraph level linguistic technology to identify substituted entities and their attributes.

References

KP Krishna Kumar, G Geethakumari. Detecting disinformation in online social networks using cognitive psychology. Human-centric Computing and Information Sciences

September 2014, 4:14.

Eons Seo, Prasant Mohapatra and Tarek Abdelzaher. Identifying Rumors and Their Sources in Social Networks. SPIE 2012.

Canini, K. R., Suh, B., and Pirolli, P. L., "Finding credible information sources in social networks based on content and social structure," in Proceedings of the 2011 IEEE Second International Conference on Social Computing, SocialCom '11, 1–8 (2011).

Mintz, Anne. "The Disinformation Superhighway?". PBS. Retrieved 26 February 2013.

Stahl, Bernd (2006). "On the Difference of Equality of Information, Disinformation, and Disinformation: A Critical Research Perspective". Informing Science 9: 83–96.

Libicki, Martin (2007). Conquest in Cyberspace: National Security and Information Warfare. New York: Cambridge University Press. pp. 51–55.

Christopher Murphy (2005). Competitive Intelligence: Gathering, Analysing And Putting It to Work. Gower Publishing, Ltd.. pp. 186–189.

Galitsky, B., Josep Lluis de la Rosa, Gábor Dobrocsi. 2012. Inferring the semantic properties of sentences by mining syntactic parse trees. Data & Knowledge Engineering. Volume 81-82, November 21-45.

Galitsky, B. 2014. Transfer learning of syntactic structures for building taxonomies for search engines. Engineering Application of AI, http://dx.doi.org/10.1016/j.engappai.2013.08.010.

Galitsky, B. 2013. Machine Learning of Syntactic Parse Trees for Search and Classification of Text. Engineering Application of AI. Volume 26, Issue 3, 1072–1091.

Boris Galitsky, Sergei O. Kuznetsov, A Web Mining Tool for Assistance with Creative Writing. 35th European Conference on Information Retrieval (ECIR 2013).

Vahed Qazvinian Emily Rosengren Dragomir R. Radev Qiaozhu Mei. Rumor has it: Identifying Misinformation in Microblogs. EMNLP-2011.

Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Goncalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2010. Detecting and trackingthe spread of astroturf memes in microblog streams. CoRR, abs/1011.3768.1599.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the newscycle. In KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 497–506.