Exploratory study of relationships among statement credibility, context, and semantic similarity

Emilia Rejmund*, Wojciech Jaworski*† and Adam Wierzbicki*

*Polish-Japanese Institute of Information Technology,
Koszykowa 86, 02-008 Warsaw, Poland (emilia.rejmund,adamw)@pjwstk.edu.pl
†Institute of Informatics, University of Warsaw,
Banacha 2, 02-097 Warsaw, Poland wjaworski@mimuw.edu.pl

***Abstract*—In our work we investigate the relationship between semantic textual similarity and credibility of the individual sentence.**

For this purpose we performed an experiment to create a corpus of sentences with known credibility value. Then we calculated semantic similarity of each pair of sentences and tried to assess the credibility of a sentence using similar one.

Performing comparison, we use idea to treat credibility as a probability distribution. By comparing the expected value of the credibility distribution with the results of the experiment, we evaluate the relationship of similarity relation with credibility.

I. Introduction

For many people Internet is a first source of information, even on so important subjects as health or finance. Unfortunately this fact is utilized to manipulate people into unchecked medical treatments, dangerous dietetic supplements, financial frauds etc. Hence, assessing the credibility of information from the Internet is a significant problem.

In this paper we propose an approach to this problem which is based on the observation that one web page may contain both, credible and non credible information. Therefore we introduce credibility assessment on the sentence level.

Since vast amount of information present in the Internet makes it impossible to evaluate every single sentence, we explore the possibility of estimating sentence credibility using credibility of sentences that are similar to it. The idea is to create a corpus of important credible and non-credible claims for a given domain and assess the credibility of sentences from that domain using the similarity relation.

There were several similarity measures proposed in the literature [6] , [7], [4] which base on surface-level and content features. However we decided to adopt the concept of Semantic Textual Similarity introduced on SemEval 2012 conference [1]. We explored the possibility of application of an algorithm basing on DkPro Similarity engine [2]. Unfortunately, we found out that such an approach is not well suited for our purpose. The reason was that it takes to much computational resources and time for our purposes. We needed a measure which is fast and transparent. Therefore, we introduced our own measure which is a simple bag of word approach we enhanced by using WordNet to generalize comparison of words.

However, before constructing any application for determining sentence's credibility we must check whether our assumption on usefulness of similarity relation for credibility assessment is correct. Therefore, we formulate the following set of hypotheses:

Hypothesis 1: Statement credibility depends strongly on context

Hypothesis 2: Similar sentences have similar credibility.

Hypothesis 3: Semantic similarity functions can be applied to predict statement credibility

In order to prove our hypotheses we developed a corpus of sentences taken from web pages concerning a controversial topic and we designed, and conducted an experiment in which respondents rated the credibility of these sentences. Thus, we obtained a corpus of sentences, with known credibility rating, on which, we validated our hypothesis.

We selected a *colloidal silver* domain for our research. This domain is very narrow, so it is easy to find similar sentences in various documents and also to gather a corpus of webpages that fully covers this domain. This domain is also controversial, which means in this case that various sources provide contradicting information concerning the influence of colloidal silver on human health.

The problem of statement credibility assessment was already studied in the literature.

TruthGoogles [9] is a browser plugin that scans a Web page's text and identifies phrases that are present in an external, treated as authoritative source of factchecked statements. However the system performs matching by means of simple string comparison, which results in 100% precision but very low recall.

Juffinger et al. [5] proposed a system for estimating blog credibility on the basis of verified content. Blog contents are treated as a bag of words assigned with their parts of speech. The similarity with verified information (namely news articles) is calculated. The value of similarity measure is then discretized into three-step credibility scale. Similarity measure and discretization thresholds make use of corpus dependent parameters and attribute transformations like TF-IDF.

Finally, Murakami et al. [8] described a vision of Statement Map Project, which is intended to present facts and opinions concerning given subject together with logical (agreement, conflict) and epistemic (evidence) relations between

5	The two sentences are completely equivalent, as they mean the same thing
4	The two sentences are mostly equivalent, but some unimportant details differ
3	The two sentences are roughly equivalent, but some important information differs/missing
2	The two sentences are not equivalent, but share some details.
1	The two sentences are not equivalent, but are on the same topic.
0	The two sentences are on different topics.

TABLE I. GOLD STANDARD - SIMILARITY DEFINITION SCALE USED IN SEMEVAL 2012 STS TASK

them. The system will retrieve documents related to a query from the Web and split them into statements. Then it will determine whether statements are fact or opinions and find relations between them. The rest of the paper is organized as follows: in Section II we describe the concept of Semantic Textual Similarity. In Section III, we introduce the notion of credibility which we employ in our experiment and in Section IV, we describe the data and the experiment design. Finally, we present experiment results and analyze them in order to validate our hypotheses in Section V.

II. SEMANTIC TEXTUAL SIMILARITY

SemEval Semantic Textual Similarity (STS) task measures the degree of semantic equivalence [1]. The goal of the task was to create an unified framework that allows for an evaluation of multiple semantic components.

We base definition of our similarity measure between two sentences on SemEval 2012 STS pilot task. Participants aim was to evaluate similarity between two sentences, resulting in a similarity score which range from 0 (no relation) to 5 (semantic equivalence). The construction of this scale is described in detail on SemEval 2012 STS (17) task [1], and we presented it on Table I.

To find a semantic textual similarity value for every pair of sentences we use the following algorithm. In the pre-processing stage each sentence is divided into individual words, then words are lemmatized and for each verb and noun all its WordNet hyperonims are added to sentence's bag of words. TF-IDF value is used to filter out most common words. Finally, similarity value is calculated as number of matching words in a bags to total number of words and transformed into a score on 1 to 5 scale according to gold standard table.

III. CREDIBILITY

In our research, we analyze the credibility rated on a 5 point Likert scale, where 5 means highly credible and 1 means highly non credible. Due to the fact that, for a given information, value of its credibility rating depends on beliefs of the person who rates, we approximate credibility by means of rating distributions:

$$cr_x = \frac{\sum_{u \in U} \chi(u,x) r(u,x)}{\sum_{u \in U} \chi(u,x)}$$

In the above formulae: U is the set of respondents that rate sentences, $r(u,x) \in \{1,\ldots,5\}$ is a value of a rating of a sentence x given by respondent u and $\chi(u,x) \in \{0,1\}$ is equal to 1 iff respondent u provided a rating for sentence x.

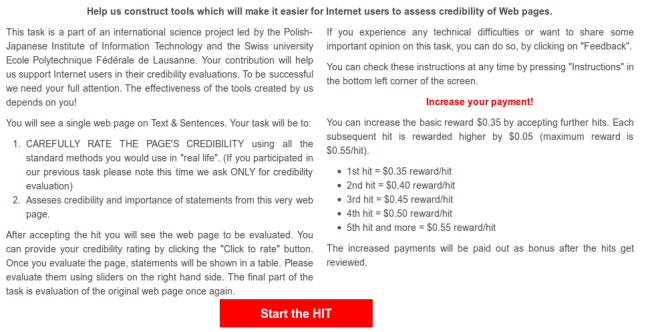

Fig. 1. Experiment instructions

IV. EXPERIMENT DESIGN

To test our hypotheses and get training data for credibility classifiers we performed an experiment in which we obtained a corpus of human rated sentences. We use this corpus to find the dependency between sentences credibility score and sentences similarity score in the SemEval STS meaning.

Experiment consists of three stages named A, B and C. In this paper we analyze only results from stage B. However we will describe briefly all the stages.

Fig. 1 contains the instruction presented to users on the beginning of an experiment. During the stage A, we showed to respondent a webpage and asked him/her for its credibility. Then, in the stage B, we showed to respondent a table with sentences extracted from the text and ask him/her questions concerning credibility and importance of these sentences. We manually chose 30 most significant sentences for each web page. Sentences in the table were presented in the same order as they appear in the text. Finally, during the third stage C, we again show a webpage to the respondent and ask him/her for its credibility. The goal of this stage is to check whether respondent's opinion about the Web page has changed after he rated singular sentences.

In all stages we used 5 point credibility scale and a color slider evaluation method . In the stage B besides a slider we added a checkbox to indicate that credibility measure is not applicable for this sentence (titles, exclamations and questions are examples of such sentences). We also asked respondent to rate the importance of each sentence with respect to the importance of the information that it contains. Finally, we asked whether the Web page contains important information that is not presented in the sentence table.

In order to eliminate additional factors, we removed hiperlinks from texts. The reason for such action is the prominence-interpretation theory [3].

We selected for the experiment 150 web pages, that fully cover all topics present in this domain. Each site was rated 19 times.

V. ANALYSIS & RESULTS

For each sentence selected for the experiment we obtained 19 ratings. We present distribution of these ratings in Fig. 2. The overwhelming amount of sentences has high credibility

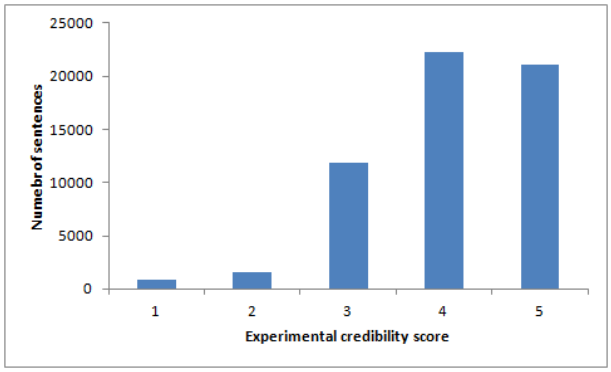

Fig. 2. Distribution of sentences credibility scores

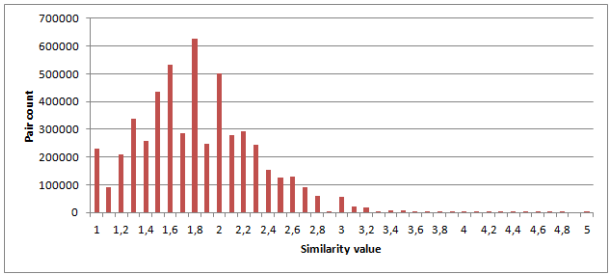

Fig. 3. Distribution of pair of sentences similarity scores for all sites

value, which is a result of the fact that even in completely non credible domain many facts is credible. In case of *colloidal silver* domain facts concerning the process of production, historical notes and even side effects are credible and non-controversial. What is controversial are the healing properties of silver. However, respondent's opinion varies even on this topic.

Next step is to calculate the value of similarity for pairs of sentences. We present its distribution on the Fig. 3. We can observe that similarity distribution concentrate around 1.8 We consider it as a result of manual selection of most important sentences and the fact that the domain of *colloidal silver* is very narrow: two randomly chosen important sentences share the same topic but are not equivalent.

We identified 583 pair of sentences with similarity 5 on different pages. We checked theirs pair credibility score, understood as an absolute value of difference between each sentence credibility value. We treat credibility score as an identical when it's equal or less then 0,5. Then we calculated the ratio of the sentences with different credibility and result about 35% shows that majority of the similar sentences from different pages has equal credibility which leads us to conclusion that credibility of the sentence does not depend on context. To prove it we compare the distribution of pair credibility value for semantically identical sentences with different context using Mann-Whitney test. The result is that they are the same with 5% significance level which false our Hypothesis 1.

Now, we will look for a connection between similarity and credibility. We split the set of sentence pairs into clusters

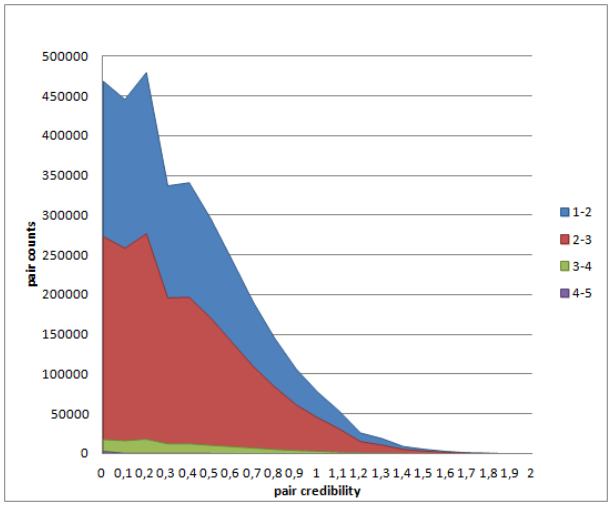

Fig. 4. Distribution of absolute value of credibility difference scores for pair of sentence for different similarity intervals

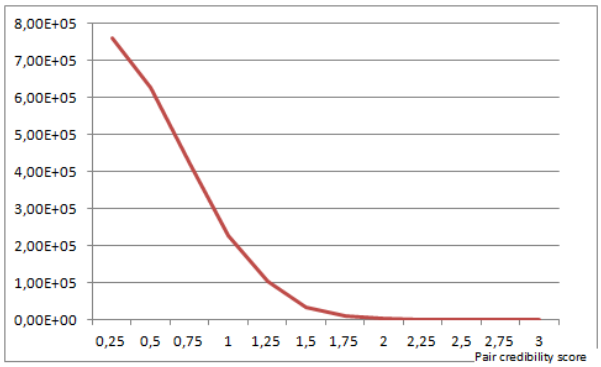

Fig. 5. Distribution of absolute value of pair of sentences credibility difference scores for all sites

according to the similarity value, then we separately for each sentence pair in cluster we calculate absolute value of difference of credibility scores. On Fig. 4 we present the distribution of absolute value of credibility difference scores sentence pairs.

The credibility difference between two sentences does not depend on it's similarity value. Number of counts which can be observed on graph is results of the construction of our corpus. This leads us to conclusion that similar sentences does not have similar credibility value, which false our Hypothesis 2.

Our goal is to determine one value of the credibility score for pair of sentence. We decide that for our purposes the best will be the absolute value of the difference of credibility for both sentences and present distribution on Fig.5. We looking for relation between credibility value for statements in pair.

To proof Hypothesis 3 we consider relation between pair of statements similarity and credibility where credibility score is understood as the absolute value of the difference for individual of sentences forming pair. Graph showing the dependence

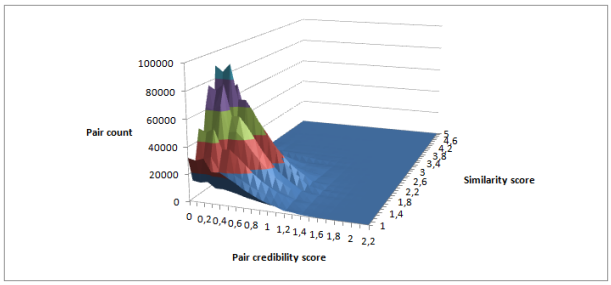

Fig. 6. Similarity and credibility value relation for every pair of sentences

s1 / s2	HC		HNC	
	similar	not similar	similar	not similar
HC	0.2102	0.0692	0.0461	0.0740
HNC	0.0430	0.0737	0.3857	0.0694

Fig. 7. Probablity of the accurate, based on similarity, prediction of credibility value for HC and HNC classes

between similarity and credibility value for pair of sentences is presented on Fig. 6

To proof our Hypothesis 3 we create a probability density function for different similarity range. The results is presented on Fig. 8. Additionally we calculate a set of conditional probability of prediction values, where, s1 and s2 denotes sentences 1 and 2, HC means highly credible, HNC means highly non credible. Treshold for HC class is 4,2 and accordingly sentence is denoted HNC when it's mean credibility score is below 3,1. The HC and NHC classes were defined by set threshold, choose the way that both classes were equinumerous. There is a third class, neutral exist but we do not consider it in our analysis because it is irrelevant for our purposes. Calculated values are presented in table on Fig. 7.

It is easy to observe, that for similar sentences value of probability that predicted credibility value is accurate, is an order of magnitude grater then for non similar statements. Applied semantic textual similarity algorithm gives best results for credible and non credible sentences. The results proof our Hypothesis 3 that semantic textual similarity preserves credibility values thus can be applied to prediction sentence

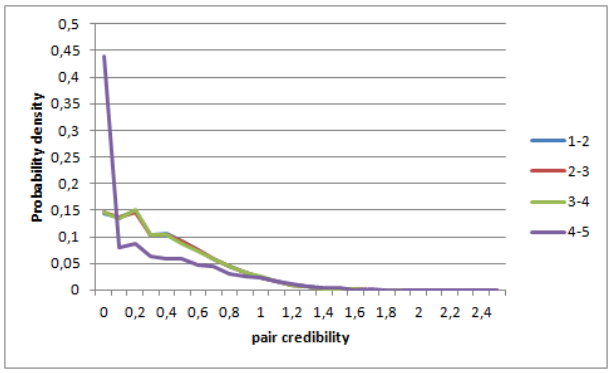

Fig. 8. Prediction probability density function

credibility based on credibility value of similar sentence.

VI. Conclusions

We formulated set of three hypotheses binding semantic textual similarity with credibility. We performed an experiment and during analysis process we proof that credibility of sentence does not depend on context. This confirms the validity of the approach treating sentence as an atomic part of web page and assessing credibility for statement not for whole page. We also proof that similar sentences doesn't have similar credibility. Finally the most interesting conclusion from our work is that similarity can predict credibility value. It opens a field of further evaluations which similarity measures fits the best to this problem.

VII. Acknowledgments

This work was financially supported by the European Community from the European Social Fund within the *INTERKADRA* project.

It was also supported by the grant *Reconcile: Robust Online Credibility Evaluation of Web Content* from Switzerland through the Swiss Contribution to the enlarged European Union.

References

[1] E. Agirre, M. Diab, D. Cer, and A. Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, SemEval '12, pages 385–393, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[2] D. Bär, C. Biemann, I. Gurevych, and T. Zesch. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, SemEval '12, pages 435–440, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[3] B. Fogg. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 722–723. ACM, 2003.

[4] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[5] A. Juffinger, M. Granitzer, and E. Lex. Blog credibility ranking by exploiting verified content. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 51–58. ACM, 2009.

[6] T. Landauer, P. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(1):259–284, 1998.

[7] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *IN AAAI'06*, pages 775–780, 2006.

[8] K. Murakami, E. Nichols, S. Matsuyoshi, A. Sumida, S. Masuda, K. Inui, and Y. Matumoto. Statement map: assisting information crediblity analysis by visualizing arguments. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 43–50. ACM, 2009.

[9] D. Schultz. *Truth goggles: automatic incorporation of context and primary source for a critical media experience*. PhD thesis, 2012.