Pilot Experiments of Hypothesis Validation Through Evidence Detection for Historians

Chris Stahlhut†‡, Christian Stab†, Iryna Gurevych†
†Ubiquitous Knowledge Processing Lab
‡ Research Training Group KRITIS
Darmstadt, Germany
www.ukp.tu-darmstadt.de

ABSTRACT

Historians spend a large amount of time in archives reading documents to pick out the small text quote they can use as evidence. This is a time consuming task that automated evidence detection promises to speed up significantly. However, no evidence detection method has been tested on a dataset that contains hypotheses and evidence created by humanities researchers. Furthermore, no research has yet been conducted to understand how historians approach this task of developing hypotheses and finding evidence. In this paper, we analyse the behaviour of 16 students of the humanities in developing and validating hypotheses and show that there is no canonical user; even when given the same exercise, they develop different hypotheses and annotate different text snippets as evidence; and current state-of-the-art argument mining methods are not suitable for historical validation of hypotheses. We therefore conclude that an evidence detection method must be trained interactively to adapt to the user's needs.

KEYWORDS

argument mining, evidence detection, hypothesis validation, information retrieval

1 INTRODUCTION

Research in humanities involves searching relevant information in huge text collections. Say, a historian analyses the political discourse after the Chernobyl and Fukushima catastrophes because he or she is working on a project about the economic development of the energy infrastructure in the second half of the 20th century. He or she will spend countless hours carefully studying protocols of political speeches and other documents, most of which do not contain any relevant information. While reading the transcript of a particular speech, the historian formulates the *hypothesis* "Extending the runtime of nuclear reactors is a monetary source of income[1]". This figurative historian then goes back to the text he or she read previously to pick out the text snippets, or *evidence*, that lead him or her to formulate the hypothesis. For instance, the statement "A depreciated nuclear reactor that runs one day longer, brings a profit of 1 million Euros." Afterwards, the historian then continues to look for more evidence supporting or attacking the hypothesis in the vast number of documents of the political discourse and, if necessary, revises the hypothesis.

In this example, the historian can benefit greatly from document retrieval as it would dramatically reduce the number of irrelevant documents to read. However, this will only help to create the bibliography of the source material which can still be very long. Picking out the few pieces of evidence contained even within this reduced number of documents still takes a lot of time. This task of finding textual sources, or *evidence*, relevant to a given hypothesis or claim is researched under the name of Evidence Detection (ED).

While ED is extensively studied in the research field of Argument Mining (AM) [6, 10], all existing methods are trained once on a fixed set of training examples; and rarely an approach focusses on researchers in the humanities as users, let alone how they develop and validate their hypotheses. Moreover, hypotheses might change over time, providing an additional challenge for static models.

In this paper, we present for the first time (1) an analysis on how scholars in the humanities develop and validate their hypotheses, (2) an analysis on the agreement of the evidence annotated by the scholars, and (3) the results of applying a state-of-the-art argument mining model for ED in the context of humanities research.

2 RELATED WORK

Existing approaches in ED focus on finding pieces of evidence to support a claim and classify their type, e.g. as statistics, expert opinions, or anecdotal evidence. This can be done to find evidence that supports a claim [6, 10] or to analyse the evidence used in online debates [1].

AM can be separated into two different approaches, namely discourse level AM and information seeking AM. The former detects arguments inside the document structure, e.g. persuasive essays [13]. The latter detects arguments depending on the predefined context [8], e.g. in the case of ED which hypothesis a piece of evidence is related to.

Fact checking [15] is a related field of growing interest in research. Its goal is to find factual evidence for or against testable statements, for instance on historical events in high school student tests [7]. Neither of these approaches allow for personalisation or focus on researchers in the humanities as users.

One area of focus in information retrieval is on supporting academic work, e.g. by finding related academic literature [5], discovering new [12], and recommending literature [4]. While supporting academics in finding documents, neither of these approaches considers the evidence contained with the relevant documents.

[1] All examples were formulated by participants of a user study in German and translated to English by us.

DESIRES 2018, August 2018, Bertinoro, Italy

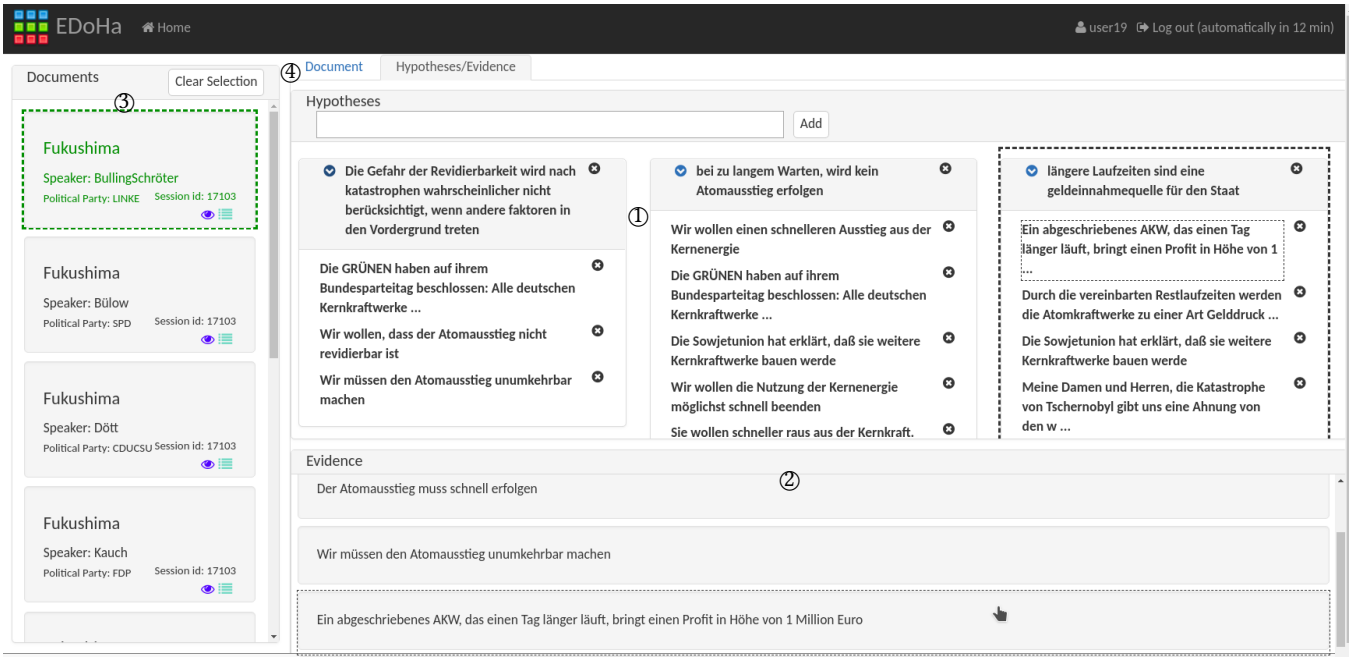

Figure 1: Screenshot of EDoHa. The Hypotheses/Evidences view allows a user to define hypotheses and link previously annotated evidence to it.

Existing approaches that work on a sub-document resolution are limited to supporting corpus exploration, for instance, by showing how the relevance of topics changes over time [11].

3 EDOHA

We developed EDoHa (Evidence Detection fOr Hypothesis vAlidation) with the goal of enabling a user to validate their hypotheses with evidence they annotated in a collection of documents. Figure 1 shows a screenshot of EDoHa in which a user already defined several hypotheses and created multiple links between hypotheses and evidence. We based EDoHa on the annotation tool WebAnno [2] but developed a user interface which focusses more on casual than expert users. It consists of the following components:

① The Hypothesis/Evidence view allows the user to define and revise hypotheses. In this screenshot, it shows three hypotheses next to each other and the evidence annotations linked to them. The hypothesis is the header and each evidence annotation linked to it is one multi-row cell beneath. Clicking the ⊗ next an evidence annotation deletes the link between the evidence annotation and the hypothesis; clicking the ⊗ next to the hypothesis deletes the hypothesis.

② A list with all evidence annotations, each of which the user can link to one or more hypotheses via Drag & Drop. To avoid showing too much evidence so that the user needs to search for them, the list of evidence annotations limits its elements to the ones from the currently selected document. When the user selects another document, the evidence annotations are replaced with the ones from the newly selected document.

③ A list of available documents in which users can annotate the evidence. The currently selected document will be shown in green colour to signify its selection. If the user wishes to see the evidence annotations from all documents, the button "Clear Selection" at the top right corner of the document list unselects the current document so that the list of evidence annotations is no longer limited to a single document. The visible hypotheses and their linked evidence are unaffected by this change.

④ A Document view in which a user can select the evidence in the source documents (not visible in the screenshot).

Interviews with historians during development showed that they require to see from which document a particular piece of evidence originates. We therefore added a highlighting mechanism to the list of available documents and Hypotheses/Evidence view. If a user hovers the cursor over an evidence, as illustrated at the bottom of the screenshot, the document source of the evidence and all hypotheses this piece of evidence is linked to will be highlighted with a dashed frame. The currently selected document will be highlighted with a green frame.

4 USER STUDY

To understand how researchers in the humanities develop and validate their hypotheses and how well they agree on the evidence, we conducted a user study with students of the humanities.

4.1 Setup

We conducted the user study in the context of a historical seminar on environmental catastrophes in the second half of the 20th century. The participants of this seminar were students of history, political science, or sociology in the second or third year of their bachelor studies. The seminar covered different historic events, such as the Chernobyl meltdown and topics of modern history, such as Waldsterben. The study took place one week after a student's presentation on the Chernobyl meltdown.

The students were asked to compare the argumentation on nuclear energy after the Chernobyl meltdown with the argumentation after the Fukushima catastrophe. We prepared 9 political speeches from the German parliament with an overall length of 479 sentences, 4 after the Chernobyl meltdown and 5 after the Fukushima catastrophe, for the students to analyse, formulate hypotheses, and validate them. The students were able to read all speeches one week beforehand to familiarise themselves with the texts. However, we did not disclose the task of the exercise to them.

Before letting the students work on the task, we gave a short introduction into the usage of EDoHa. Afterwards, we handed out the exercise and answered all questions the students had regarding it.[2] The students had one hour for the exercise followed by filling out a questionnaire about their approach to evidence detection and hypothesis validation, whether or not they would like to use EDoHa in their studies, and how to improve. The session ended with a discussion of the student's findings.

During the experiment, we logged multiple interactions of the users with the system to understand how they develop and validate hypotheses. These interactions are: *clicking on a document in the list of available documents, creating and deleting evidence annotations in documents, creating and deleting evidence/hypothesis links, creating and updating hypotheses* (reformulating or deleting hypotheses), and *changing the view in the interface.*

4.2 User behaviour

We used the previously described logs to understand their general approach to developing and validating hypotheses. Figure 2 shows the variability of how users annotate evidence and link them to hypotheses.

The upper six plots show a strong separation into distinct phases of *evidence collection* and *hypotheses validation*, or a *phased approach*. The first two users never reach the *hypothesis validation* phase, but the following four always start by collecting multiple evidence annotations and then linking them to one or more hypotheses. Afterwards, they continue to collect more evidence.

At the bottom we see that user19 and user17 showed no such distinction, i.e. these users used a *phase-free approach*. They create one evidence annotation and link it immediately to a hypothesis. Afterwards they create the next evidence annotation and link this one. In the middle, we see a transition from users with a *phased approach* towards a *phase-free approach*.

[2]The exercise sheet also contained the login credentials of previously created accounts (user0 – user20) and cannot be traced back to individual students. It is our understanding of the regulations at our institution that an ethics approval is only required when processing personally identifiable information. Being aware of the delicate nature of such data, we decided to not collect any personally identifiable information and designed the study to be anonymised as described above.

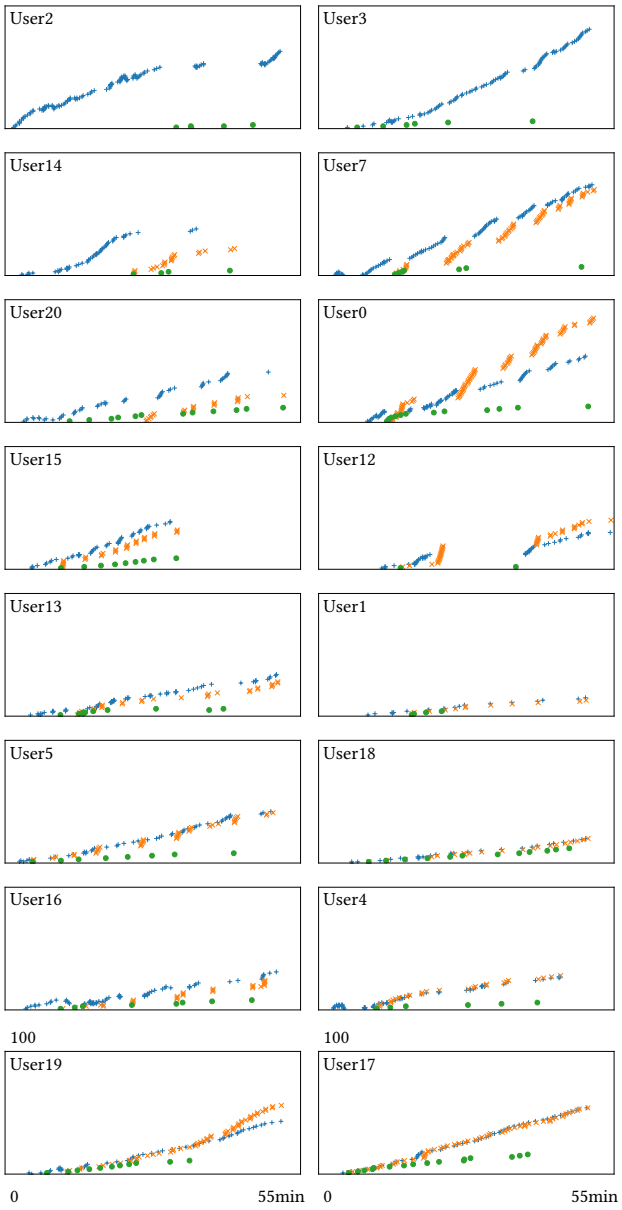

Figure 2: The number of evidence annotations (+), evidence/hypothesis links (x) for validation, and hypotheses (•) over time. The users are ordered by the number of times they changed between the Document and Hypotheses/Evidence view.

The user's approach towards which hypotheses they validated also fell into two categories. Figure 3 shows each hypothesis as a layer where its thickness represents how much evidence is linked to it. About half of the users validated multiple hypotheses at the same time, or *concurrently* (figure 3 left), whereas the other half validated the hypotheses *sequentially*, creating links to evidence for one hypothesis at a time, never returning to it (figure 3 right). We

Table 1: Agreement on evidence of similar hypotheses (top) and all hypotheses with a substantial agreement (bottom).

Hypotheses pair		Cohen's κ
International security arrangements in the nuclear sector are necessary	Nuclear power and security: further expansion of domestic and foreign policy	0.116
Nuclear-phaseout is not possible due to the profit motive of corporations	Profit maximisation of the economy	0.067
Chernobyl as a reminder for the nuclear-phaseout	Chernobyl and Fukushima repeatedly related	0.057
Nuclear phase-out should not be slowed down by individual companies	Is money and the economy put on the safety of each one?	-0.007
Does the nuclear industry have too much power?	Criticism of Fukushima	1.000
Security of nuclear reactors must be guaranteed	The following security measurements	0.748
Does the nuclear industry have too much power?	If the information policy comes from one actor, there is a high probability that not all information will reach the public	0.666
Criticism of Fukushima	If the information policy comes from one actor, there is a high probability that not all information will reach the public	0.666

found no connection between validating hypotheses sequentially and using discrete phases for evidence collection and hypothesis validation, e.g. users of a phased-approach also worked concurrently on multiple hypotheses.

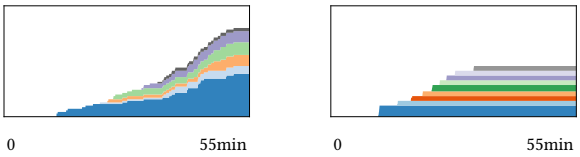

Figure 3: About half of the users validated multiple hypotheses at the same time (left), while the others validated only one hypothesis at a time, not getting back afterwards (right).

Most users (11 of 16) reported that they collected evidence first, and formulated their hypotheses later. However, while almost all users did start with the evidence collection task, many of them formulated hypotheses very early in the task and linked evidence to them at a later time, resembling a mixed approach. Only one user reported to have used a mixed approach of collecting evidence and defining hypotheses.

The behaviour of the users shows a great variety in how they develop and validate hypotheses. We also found that our current user interface does not support the phase-free approach very well. A user following the phase-free approach has to switch from the Document view to the Hypothesis/Evidence view and back to link the just created evidence annotation to a hypothesis and collect the next piece of evidence.

4.3 Agreement of the users on evidence

We followed two approaches to understand how well the users agreed on the evidence: (1) how well do the users agree on evidence for similar hypotheses and (2) how similar are the hypotheses whose evidence shows a substantial agreement.

We calculated the agreement of pairs of hypotheses (h_1, h_2) by first, creating two copies of the un-annotated documents, one for each hypothesis; and second, annotating in the first copy only the sentences that were annotated as evidence and linked to h_1 and in the second copy the ones that were linked to h_2. We then calculated Cohen's κ on these sentential annotations of the two copies.

To understand the agreement on similar hypotheses, we asked a historian to select closely related pairs of hypotheses. The agreement in visible in table 1 at the top.

In our second approach, we calculated the agreement of all pairs of hypotheses from different users. This left us with 6050 hypotheses pairs. The bottom of table 1 shows all hypotheses pairs from different users that show a substantial agreement of $\kappa > 0.6$.

Our results show that users who formulate similar hypotheses do not agree on the evidence and the same evidence can be used to validate vastly different hypotheses. This means that to maximise its usefulness, e.g. by avoiding to suggest uninteresting pieces of evidence, an ED method must adapt to the user.

4.4 Statistics on the evidence and hypotheses created by the users

In the user study, we collected 827 evidence annotations, 114 unique hypotheses (two users formulated two identical hypotheses), and 516 links between evidence annotations and hypotheses[3]. Table 2 breaks the collected data down into the number sentences in the documents the user opened, evidence annotations, hypotheses, overall, and the average number of links between hypotheses and evidence for each user.

The variability of the collected data mirrors the differences in the user's behaviour. Some users created few annotations, whereas others created many. Equally variable is the number of hypotheses and links between hypotheses and evidence. For instance, user12 created only two hypotheses, one with 11 links and the other one

[3]We plan to publish EDoHa and the data together with a more detailed evaluation of ED methods. Until then, the data is available upon request.

Table 2: The number of evidence annotations, hypotheses, and links between them varied greatly between users.

User	Sentences	Evidence	Hypotheses	Links
user0	364	205	13	259
user1	321	21	4	12
user2	403	79	3	0
user3	479	85	6	0
user4	479	27	6	27
user5	479	78	8	63
user7	479	74	7	70
user12	479	29	2	29
user13	403	38	6	30
user14	479	41	4	23
user15	479	38	9	32
user16	441	41	8	28
user17	321	77	16	61
user18	291	45	12	44
user19	479	44	11	56
user20	328	38	12	21

with 18. User18 on the other hand created 12 hypotheses and linked them with up to three evidence annotations. However, users who created many hypotheses did not always create fewer links between evidence and hypothesis than users who created few hypotheses, as user7 demonstrates with 7 hypotheses and an average of 10 evidence links. Users 2 and 3 did not create any links between evidence and hypotheses. Interactions with the participants during the study led us to believe that user2 did not understand the purpose of the study and treated it as a usability test in which the hypotheses and evidence could not be connected. User3 may have missed the linking part of the introduction into EDoHa and may therefore have been unaware of the Drag & Drop functionality.

5 EVIDENCE DETECTION EXPERIMENTS

We treated ED as a binary classification task, *evidence* vs. *no evidence*, on the sentence level and report the standard metrics (precision, recall, and F1-score). We are especially interested in the precision on the evidence class, because suggesting pieces of evidence that the user is not interested in means additional work for corrections, thereby reducing the acceptance of the system. When reporting the results on both classes, evidence and no evidence, we caclulated the macro-average precision and recall and macro-averaged F1-score from them.

We evaluated multiple baselines, models trained on the data of individual users, pre-trained models, and combinations of pre-trained models with filters that were derived from the user-created data.

Based on our previous finding that each user requires a unique ED model, we ran the experiments for each user separately. We conducted the experiments in a leave-one-document-out fashion, i.e. in each fold we used one document for testing and the others as training documents; we ignored documents the user did not open. When evaluating a non-deterministic model, e.g. neural networks or a random baseline, we repeated the experiment five times and averaged the results.

All hyperparameter optimisations were done on a development user. We chose user7 because this user annotated much evidence, created multiple hypotheses, and validated them well; methods that wouldn't work for this user because they require more data would also not work for all the others.

5.1 Baselines and models trained on user-created data

As baseline methods, we chose a majority classifier and a random classifier that learns the distribution of the training labels and predicts randomly according to them. Additionally, we trained a Multi-Layer Perceptron (MLP) with one hidden layer of size 10 and a Naive Bayes classifier. Both models rely on a bag of words as features. The MLP and Naive Bayes classifiers were implemented using scikit-learn[4] and stopwords were removed based on NLTK[5].

We also considered the links between evidence and hypotheses as training data. This classifier (link(s, h)) was trained to predict the link between hypotheses and evidence. The negative samples for training were random links between evidence and hypotheses, and positive samples were the user-created links. We used an MLP with three hidden layers (100, 75, and 50 nodes) to predict the binary link between evidence and hypotheses. It used averaged word embeddings in German trained on articles from the newspaper "Die Zeit" for a GermEval 2014 task on nested named entity recognition [9]. If this classifier detected a link between a sentence and a user-defined hypothesis, it considered the sentence a piece of evidence.

5.2 Pre-trained models for argument mining

As AM model, we selected a bidirectional Long-Short Term Memory that uses a candidate sentence and the cosine similarity between the candidate and the topic as input. We trained it on the sentential AM corpus created by Stab et. al [14], limiting the data to the topic of nuclear energy. To adapt the model to the German language, we translated the sentences into German using an external machine translation API[6] similar to [3]. The model reached a macro F1-score of 0.714 in a binary in-topic classification task of argument vs. no argument. In our ED task, we treated sentences which the model classified as argumentative as evidence.

5.3 User data augmented models

To investigate whether the user-created data can be used to augment a pre-trained model, we developed three approaches that combined the user-created data with the best performing pre-trained model. We used the following methods to reduce the number of false evidence suggestions by filtering the predictions of the pre-trained model with:

+cos(h, s) Cosine similarity between hypothesis and predicted evidence < 0.7.

+ignore < 60s A heuristic that ignores all predictions on files the user did not open for at least 60s, because a user may not spend much time reading documents that are deemed irrelevant.

+link(h, s) Prediction of a link between the evidence predicted by the pre-trained model and any hypothesis the user created.

[4] https://scikit-learn.org/stable/
[5] https://www.nltk.org/
[6] We chose the Google Translate API because of the quality of the translations.

Table 3: Results in the ED task were averaged across users with standard deviation in parentheses. The bottom shows combinations of the best performing model with additional user generated data. A † indicates a statistically significant difference to the random baseline and ‡ indicates a statistically significant difference to the AM model. Both significances are calculated across users using a Wilcoxon signed rank test with Pratts' modification with zero rank splitting and a threshold of $p < 0.05$.

	Evidence & No Evidence			Evidence only		
	Macro F1	Macro P	Macro R	F1	P	R
Majority	† 0.462 (0.032)	† 0.433 (0.049)	† 0.500 (0.000)	† 0.051 (0.186)	† 0.040 (0.145)	† 0.071 (0.258)
Random	0.491 (0.013)	0.491 (0.012)	0.491 (0.013)	0.126 (0.132)	0.127 (0.131)	0.126 (0.133)
MLP	† 0.526 (0.037)	† 0.538 (0.060)	† 0.516 (0.019)	† 0.132 (0.138)	† 0.213 (0.158)	† 0.104 (0.129)
NaiveBayes	0.506 (0.029)	0.506 (0.024)	0.507 (0.035)	0.169 (0.123)	0.151 (0.139)	0.202 (0.119)
cos(h, s)	† 0.506 (0.143)	† 0.505 (0.144)	† 0.508 (0.145)	† 0.217 (0.169)	† 0.152 (0.145)	† **0.768 (0.328)**
link(h, s)	0.441 (0.136)	0.445 (0.140)	0.444 (0.141)	0.123 (0.101)	0.091 (0.075)	0.300 (0.238)
AM	† **0.574 (0.020)**	† **0.548 (0.024)**	† **0.604 (0.026)**	† **0.265 (0.101)**	† **0.208 (0.154)**	† 0.511 (0.059)
AM+cos(h, s)	‡ 0.489 (0.139)	‡ 0.476 (0.133)	‡ 0.502 (0.146)	‡ 0.206 (0.119)	‡ 0.146 (0.134)	‡ **0.516 (0.158)**
AM+ignore < 60s	**0.585 (0.037)**	**0.560 (0.041)**	**0.613 (0.040)**	**0.282 (0.113)**	**0.230 (0.164)**	0.490 (0.077)
AM+link(h, s)	‡ 0.450 (0.129)	‡ 0.454 (0.129)	‡ 0.446 (0.131)	‡ 0.182 (0.129)	‡ 0.124 (0.119)	‡ 0.492 (0.154)

5.4 Results

Table 3 shows that the AM model performs best among the baselines. However, the differences between it and the Random baseline were generally not statistically significant.

Among the models trained on the user data, the MLP outperformed all others with respect to precision on the evidence class followed by the Naive Bayes and cos(h, s). The cos(h, s) model reached the overall highest recall on the evidence class, but did so at the cost of predicting many false positives.

The link(h, s) model performed unexpectedly low, which can be due to the nature of the training data. Because the training data consisted of positive links between evidence and hypotheses, the negative samples were drawn from existing evidence, just paired with a random hypothesis. The training data therefore did not contain any sentence that the user did not annotate as evidence, leading the model to predict greetings as evidence.

Contrary to our initial assumption, some users did create evidence annotations in documents that they opened for less than the 60s. This resulted in the drop in recall between the AM and AM+ignore < 60s model. Nevertheless, ignoring the files that the user did not open for more than 60s did improve the performance significantly in any measure except the recall on the evidence class.

Overall, no method achieved sufficient results meaning that their integration into an ED tool is not yet feasible. Especially the low precision on the evidence class would discourage any adoption. However, the performance of the pre-trained AM model is promising regarding further training to adapt an ED model to individual users.

6 CONCLUSION

In this paper we have presented the first prototype of an evidence detection and hypothesis validation tool developed with humanities researchers as users in mind. We conducted a user study with bachelor students to understand how they develop and validate their hypotheses in history and found that users vary greatly when collecting evidence and validating hypotheses. We also found that even though all participants were given the same task, each of them created unique hypotheses and evidence annotations. Furthermore, similar hypotheses were not supported with the same evidence and the same evidence was used to support different hypotheses. Given that pre-training an ED for each user is infeasible we conclude that in ED for humanities researchers the model has to be trained interactively by the user. When applying a state-of-the-art AM model to the task of ED we found that it performed better than the models trained on the user's data; we therefore conclude that a pre-trained AM model can serve as a starting point for adapting an ED to individual users.

In the future, we intend to improve the ED methods, use a more realistic setup rather than leave-one-document-out, e.g. by predicting the annotations the user is going to do next, and collect evidence and hypotheses from users working with EDoHa for longer than one hour.

ACKNOWLEDGMENTS

This work has been supported by the German Research Foundation (DFG) as part of the Research Training Group KRITIS No. GRK 2222/1, by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 03VP02540 (ArgumenText), and by the German Federal Ministry of Education and Research under the promotional reference 01UG1816B (CEDIFOR).

REFERENCES

[1] Aseel Addawood and Masooda Bashir. 2016. "What Is Your Evidence?" A Study of Controversial Topics on Social Media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. Association for Computational Linguistics, Berlin, Germany, 1–11. https://doi.org/10.18653/v1/W16-2801

[2] Richard Eckart de Castilho, Eva Mujdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A Web-Based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. 76–84.

[3] Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-Lingual Argumentation Mining: Machine Translation (and a Bit of Projection) Is All You Need!. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. to appear.

[4] Stefan Feyer, Sophie Siebert, Bela Gipp, Akiko Aizawa, and Joeran Beel. 2017. Integration of the Scientific Recommender System Mr. DLib into the Reference Manager JabRef. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*. Springer, Cham, Aberdeen, Scotland UK, 770–774. https://doi.org/10.1007/978-3-319-56608-5_80

[5] Matthias Hagen, Anna Beyer, Tim Gollub, Kristof Komlossy, and Benno Stein. 2016. Supporting Scholarly Search with Keyqueries. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*. Springer, Cham, Padua, Italy, 507–520. https://doi.org/10.1007/978-3-319-30671-1_37

[6] Xinyu Hua and Lu Wang. 2017. Understanding and Detecting Supporting Arguments of Diverse Types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 203–208.

[7] Mio Kobayashi, Ai Ishii, Chikara Hoshino, Hiroshi Miyashita, and Takuya Matsuzaki. 2017. Automated Historical Fact-Checking by Passage Retrieval, Word Statistics, and Virtual Question-Answering. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 967–975.

[8] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context Dependent Claim Detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 1489–1500.

[9] Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. GermEval-2014: Nested Named Entity Recognition with Neural Networks. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, Gertrud Faaß and Josef Ruppenhofer (Eds.). Universitätsverlag Hildesheim, Hildesheim, Germany, 117–120.

[10] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 440–450.

[11] Carsten Schnober and Iryna Gurevych. 2015. Combining Topic Models for Corpus Exploration: Applying LDA for Complex Corpus Research Tasks in a Digital Humanities Project. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications (TM '15)*. ACM, New York, NY, USA, 11–20. https://doi.org/10.1145/2809936.2809939

[12] Amin Sorkhei, Kalle Ilves, and Dorota Glowacka. 2017. Exploring Scientific Literature Search Through Topic Models. In *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics (ESIDA '17)*. ACM, Limassol, Cyprus, 65–68. https://doi.org/10.1145/3038462.3038464

[13] Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics* 43, 3 (Sept. 2017), 619–659.

[14] Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-Topic Argument Mining from Heterogeneous Sources Using Attention-Based Neural Networks. *arXiv:1802.05758 [cs]* (Feb. 2018). arXiv:cs/1802.05758

[15] Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task Definition and Dataset Construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Association for Computational Linguistics, Baltimore, MD, USA, 18–22. https://doi.org/10.3115/v1/W14-2508