Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment*

Martin Potthast,[1] Steve Göring,[1] Paolo Rosso,[2] and Benno Stein[1]

[1]Web Technology & Information Systems, Bauhaus-Universität Weimar, Germany
[2]Natural Language Engineering Lab, Universitat Politècnica de València, Spain

pan@webis.de http://pan.webis.de

Abstract This paper reports on the organization of a new kind of shared task that outsources the creation of evaluation resources to its participants. We introduce the concept of data submissions for shared tasks, and we use our previous shared task on text alignment as a testbed. A total of eight evaluation datasets have been submitted by as many participating teams. To validate the submitted datasets, they have been manually peer-reviewed by the participants. Moreover, the submitted datasets have been fed to 31 text alignment approaches in order to learn about the datasets' difficulty. The text alignment implementations have been submitted to our shared task in previous years and since been kept operational on the evaluation-as-a-service platform TIRA.

1 Introduction

The term "shared task" refers to a certain kind of computer science event, where researchers working on a specific problem of interest, *the task*, convene to compare their latest algorithmic approaches at solving it in a controlled laboratory experiment.[1] The organizers of a shared task usually take care of the lab setup by compiling evaluation resources, by selecting performance measures, and sometimes even by raising the task itself for the first time.

The way shared tasks are organized at present may strongly influence future evaluations: for instance, in case the event has sufficient success within its target community, subsequent research on the task may be compelled to follow the guidelines proposed by the shared task organizers, or else risk rejection from reviewers who are aware of the shared task. The fact that a shared task has been organized may amplify the importance of its experimental setup over others, stifling contributions off the beaten track. However, there is only anecdotal evidence in support of this narrowing effect of shared tasks. Nevertheless, it has been frequently pointed out in the position papers submitted to a workshop organized by the natural language generation community on the pros and cons of adopting shared tasks for evaluation [1].

* A summary of this report has been published as part of [62], and some of the descriptions are borrowed from earlier reports on the shared task of text alignment at PAN.

[1] The etymology of the term "shared task" is unclear; conceivably, it was coined to describe a special kind of conference track and was picked up into general use from there.

One of the main contributions of organizing a shared task is that of creating a reusable experimental setup for future research, allowing for comparative evaluations even after the shared task has passed. Currently, however, this goal is only partially achieved: participants and researchers following up on a shared task may only compare their own approach to those of others, whereas other aspects of a shared task remain fixed, such as the evaluation datasets, the ground truth annotations, and the performance measures used. As time passes, these fixtures limit future research to using evaluation resources that may quickly become outdated in order to compare new approaches with the state of the art. Moreover, given that shared tasks are often organized by only a few dedicated people, this further limits the attainable diversity of evaluation resources.

To overcome the outlined shortcomings, we propose that shared task organizers attempt to remove as many fixtures from their shared tasks as possible, relinquishing control over the choice of evaluation resources to their community. We believe that, in fact, only data formats and interfaces between evaluation resources need to be fixed a priori to ensure compatibility of contributions submitted by community members. As a first step in this direction, we investigate for the first time the feasibility of data submissions to a well-known shared task by posing the construction of evaluation datasets as a shared task of its own.

As a testbed for data submissions, we use our established shared task on plagiarism detection at PAN [62], and in particular the task on text alignment. Instead of inviting text alignment algorithms, we ask participants to submit datasets of their own design, validating the submitted datasets both via peer-review and by running the the text alignment softwares submitted in previous editions of the text alignment task against the submitted corpora. Altogether, eight datasets have been submitted by participants, ranging from automatically constructed ones to manually created ones, including various languages.

In what follows, after a brief discussion of related work in Section 2, we outline our approach to data submissions in Section 3, survey the submitted datasets in Section 4, and report on their validation and evaluation in Section 5.

2 Related Work

Research on plagiarism detection has a long history, both within PAN and without. We have been the first to organize shared tasks on plagiarism detection [51], whereas since then, we have introduced a number of variations of the task as well as new evaluation resources: the first shared task was organized in 2009, studying two sub-problems of plagiarism detection, namely the traditional external plagiarism detection [64], where a reference collection is used to identify plagiarized passages, and intrinsic plagiarism detection [31, 63], where no such reference collection is at hand and plagiarism has to be identified from writing style changes within a document. For the this share task, we have created the first standardized, large-scale evaluation corpus for plagiarism detection [50]. As part of this effort, we have devised novel performance measures which, for the first time, took into account task-specific characteristics of plagiarism detection, such as detection granularity. Finally, in the first three years of PAN, we have introduced cross-language plagiarism detection as a sub-task of plagiarism detection

for the first time [42], adding corresponding problem instances into the evaluation corpus. Altogether, in the first three years of our shared task, we successfully acquired and evaluated plagiarism detection approaches of 42 research teams from around the world, some participating more than once. Many insights came out of this endeavor which informed our subsequent activities [51, 41, 43].

Starting in 2012, we have completely overhauled our evaluation approach to plagiarism detection [44]. Since then, we have separated external plagiarism detection into the two tasks source retrieval and text alignment. The former task deals with information retrieval approaches to retrieve potential sources for a suspicious document from a large text collection, such as the web, which are indexed with traditional retrieval models. The latter task of text alignment focuses on the problem of extracting matching passages from pairs of documents, if there are any. Both tasks have never been studied in this way before.

For source retrieval, we went to considerable lengths to set up a realistic evaluation environment: we indexed the entire English portion of the ClueWeb09 corpus, building the research search engine ChatNoir [48]. ChatNoir served two purposes, namely as an API for plagiarism detectors for those who cannot afford to index the ClueWeb themselves, but also as an end user search engine for authors which were hired to construct a new, realistic evaluation resource for source retrieval. We have hired 18 semi-professional authors from the crowdsourcing platform oDesk (now Upwork) and asked them to write essays of length at least 5000 words on pre-defined TREC web track topics. To write their essays, the authors were asked to conduct their research using ChatNoir, reusing text from the web pages they found. This way, we have created realistic information needs which in turn lead the authors to use our search engine in a realistic way to fulfill their task. AS part of this activity, we gained new insights into the nature of how humans reuse text, some building up a text as they go, whereas others first collect a lot of text and then boil it down to the final essay [49]. Finally, we have devised and developed new evaluation measures for source retrieval that, for the first time, take into account the retrieval of near-duplicate results when calculating precision and recall [45, 47]. We report on the latest results of the source retrieval subtask in [20].

Regarding text alignment, we focus on the text reuse aspects of the task by stripping down the problem to its very core, namely comparing two text documents to identify reused passages of text. In this task, we have started in 2012 to experiment with software submissions for the first time, which lead to the development of the TIRA experimentation platform [18]. TIRA is an implementation of the emerging evaluation-as-a-service paradigm [22]. We have since scaled TIRA in order to also collect participant software for source retrieval and for the entire PAN evaluation lab as of 2013, thus improving the reproducibility of PAN's shared tasks for the foreseeable future [17, 46]. Altogether, in the second three-year cycle of this task, we have acquired and evaluated plagiarism detection approaches of 20 research teams on source retrieval and 31 research teams on text alignment [44, 45, 47].

3 Data Submissions: Crowdsourcing Evaluation Resources

Data submissions for shared tasks have not been systematically studied until now, so that no best practices have been established, yet. Asking shared task participants to submit data is nothing short of crowdsourcing, albeit the task of creating an evaluation resource is by comparison much more complex than average crowdsourcing tasks found in the literature. In what follows, we outline the rationale of data submissions, review important aspects of defining a data submissions task that may inform instructions to be handed out to participants, and detail two methods to evaluating submitted datasets.

3.1 Rationale of Data Submissions to Shared Tasks

Traditionally, the evaluation resources required to run a shared task are created by its organizers—but the question remains: why? The following reasons can be identified:

- *Quality control.* The success of a shared task rests with the quality of its evaluation resources. A poorly built evaluation dataset may invalidate evaluation results, which is one of the risks of organizing shared tasks. This is why organizers have a vested interest in maintaining close control over evaluation resources, and how they are constructed.
- *Seniority.* Senior community members may have the best vantage point in order to create representative evaluation resources.
- *Access to proprietary data.* Having access to an otherwise closed data source (e.g., from a company) gives some community members an advantage over others in creating evaluation resources with a strong connection to the real world.
- *Task inventorship.* The inventor of a new task (i.e., tasks that have not been considered before), is in a unique position to create normative evaluation resources, shaping future evaluations.
- *Being first to the table.* The first one to pick up the opportunity may take the lead in constructing evaluation resources (e.g., when a known task has never been organized as a shared task before, or, to mitigate a lack of evaluation resources).

All of the above are good reasons for an individual or a small group of researchers to organize a shared task, and, to create corresponding evaluation resources themselves. However, from reviewing dozens of shared tasks that have been organized in the human language technologies, neither of them are a necessary requirement [46]: shared tasks are being organized using less-than-optimal datasets, by newcomers to a given research field, without involving special or proprietary data, and without inventing the task in the first place. Hence, we question the traditional connection of shared task organization and evaluation resource construction. This connection limits the scale and diversity, and therefore the representativeness of the evaluation resources that can be created:

- *Scale.* The number of man-hours that can be invested in the construction of evaluation resources is limited by the number of organizers and their personal commitment. This limits the scale of the evaluation resources. Crowdsourcing may be employed as a means to increase scale in many situations, however, this is mostly not the case when task-specific expertise is required.

– *Diversity.* The combined task-specific capabilities of all task organizers may be limited regarding the task's domain. For example, the number of languages spoken by task organizers is often fairly small, whereas true representativeness across languages would require evaluation resources from at least all major language families spoken today.

By involving participants in a structured way into the construction of evaluation resources, task organizers may build on their combined expertise, man-power, and diversity. However, there is no free lunch, and outsourcing the construction of evaluation resources introduces the new organizational problem that the datasets created and submitted by third parties must be validated and evaluated for quality.

3.2 Defining a Data Submission Task

When casting a data submission task, there are a number of desiderata that participants should meet:

– *Data format compliance.* The organizers should agree on a specific data format suitable for the task in question. The format should be defined with the utmost care, since it may be impossible to fix mistakes discovered later on. Experience shows that the format of the evaluation datasets has a major effect on how participants implement their softwares for a task, which is especially true when inviting software submissions for a shared task. Regarding data submissions, a dataset should comprise a set of problem instances with respect to the task, where each problem instance shall be formatted according to the specifications handed out by the organizers. To ensure compliance, the organizers should prepare a format validation tool, which allows participants to check the format of their dataset in progress, and whether it complies with the format specifications. This way, participants move into the right direction from the start, and less back and forth will be necessary after a dataset has been submitted. The format validation tool should check every aspect of the required data format in order to foreclose any unintended deviation.
– *Annotation validity.* All problem instances of a dataset should comprise annotations that reveal their true solution with regard to the task in question. It goes without saying, that all annotations should be valid. Datasets that do not comprise annotations are of course useless for evaluation purposes, whereas annotation validity as well as the quality and representativeness of the problem instances selected by participants determines the usefulness of a submitted dataset.
– *Representative size.* The datasets submitted should be of sufficient size, so that dividing them into training and test datasets can be done without sacrificing representativeness, and so that evaluations conducted based on the resulting test datasets are meaningful and not prone to noise.
– *Choice of data source.* The choice of a data source should be left up to participants, and should open the possibility of using manually created data either from the real world or by asking human test subjects to emulate problem instances, as well as automatically generated data based on a computer simulation of problem instances for the task at hand.

– *Copyright and sensitive data.* Participants must ensure that they have the usage rights of the data, for transferring usage rights to the organizers of the shared task, and for allowing the organizers to transfer usage rights to other participants. The data must further be compliant with privacy laws and ethically innocuous. Dependent on the task at hand and what the organizers of a shared task desire, accepting confidential or otherwise sensitive data may still be possible: in case the shared task also invites software submissions, the organizers may promise participants that the sensitive data does not leak to participants by running submitted software at their site against the submitted datasets. Nevertheless, special security precautions must be taken to ensure that sensitive data does not leak when feeding it to untrusted software.

3.3 Evaluating Submitted Datasets: Peer-Review and Software Submissions

The construction of new evaluation datasets must be done with the utmost care, since datasets are barely double-checked or questioned again once they have been accepted as authoritative. This presents the organizers of a dataset construction task with the new challenge of evaluating submitted datasets, where the evaluation of a dataset should aim at establishing its validity. In general, the organizers of data submission tasks should ensure not to advertise submitted datasets as valid unless they are, since such an endorsement may carry a lot of weight in a shared task's community.

Unlike with shared tasks that invite algorithmic contributions, the validity of a dataset typically can not be established via an automatically computed performance measure, but requires manual reviewing effort. Therefore, as part of their participation, all participants who submit a dataset should be compelled to also peer-review the datasets submitted by other participants. Moreover, inviting other community members to conduct independent reviews may ensure impartial results. Reviewers may be instructed as follows:

> The peer-review is about dataset validity, i.e. the quality and realism of the problem instances. Conducting the peer-review includes:
> – *Manual* review of as many examples as possible from all datasets
> – Make observations about how the dataset has been constructed
> – Make observations about potential quality problems or errors
> – Make observations on the realism of each dataset's problem instances
> – Write about your observations in your notebook (make sure to refer to examples from the datasets for your findings).

Handing out the complete submitted datasets for peer-review, however, is out of the question, since this would defeat the purpose of subsequent shared task evaluations by revealing the ground truth prematurely. Here, the organizers of a dataset construction task serve as mediators, splitting submitted datasets into training and test datasets, and handing out only the training portion for peer-review. The participants who submitted a given dataset, however, may never be reliably evaluated based on their own dataset. Also, colluding participants may not be ruled out entirely.

Finally, when a shared task has previously invited software submissions, this creates ample opportunity to re-evaluate the existing softwares on the submitted datasets. This

allows for evaluating submitted datasets in terms of their difficulty: the performances of existing software on submitted datasets, when compared to their respective performances on established datasets, allow for a relative assessment of dataset difficulty. If a shared task did not invite software submissions so far, then the organizers should set up a baseline software for the shared task and run that against submitted datasets to allow for a relative comparison among them.

3.4 Data Submissions for Text Alignment

In text alignment, given a pair of documents, the task is to identify all contiguous passages of reused text between them. The challenge with this task is to identify passages of text that have been obfuscated, sometimes to the extent that, apart from stop words, little lexical similarity remains between an original passage and its reused counterpart. Consequently, for task organizers, the challenge is to provide a representative dataset of documents that emulate this situation. For the previous editions of PAN, we have created such datasets ourselves, whereas obfuscated text passages have been generated automatically, semi-automatically via crowdsourcing [6], and by collecting real cases. Until now, however, we neglected participants of our shared task as potential assistants in creating evaluation resources. Given that a stable community has formed around our task in previous years, and that the data format has not changed in the past three years, we felt confident to experiment with this task and to switch from algorithm development to data submissions. We cast the task to construct an evaluation dataset as follows:

- *Dataset collection.* Gather real-world instances of text reuse or plagiarism, and annotate them.
- *Dataset generation.* Given pairs of documents, generate passages of reused or plagiarized text between them. Apply a means of obfuscation of your choosing.

The task definition is kept as open as possible, imposing no particular restrictions on the way in which participants approach this task, which languages they consider, or which kinds of obfuscation they collect or generate. In particular, the task definition highlights the two possible avenues of dataset construction, namely manual collection, and automatic construction. To ensure compatibility among each other and with previous datasets, however, the format of all submitted datasets had to conform with that of the existing datasets used in previous years. By fixing the dataset format, future editions of the text alignment task may build on the evaluation resources created within this task without further effort, and the pieces of software that have been submitted in previous editions of the text alignment task, which are available on the TIRA platform for evaluation as a service, may be re-evaluated on the new datasets. In our case, more than 30 text alignment approaches have been submitted since 2012. To ensure compatibility, we handed out a dataset validation tool that checked all format restrictions.

4 Survey of Submitted Text Alignment Datasets

A total of eight datasets have been submitted to the PAN 2015 text alignment task on dataset construction. The datasets are of varying sizes and have been built with a variety

of methods. In what follows, after a brief discussion of the dataset format, we survey the datasets with regard their source of documents and languages, and the construction process employed by their authors, paying special attention to the obfuscation approaches. The section closes with an overview of dataset statistics.

4.1 Dataset Format

We asked participants to comply with the dataset format of the PAN plagiarism corpora that have been used for the text alignment task at PAN since 2012. A compliant dataset consists of documents in the form of plain text files encoded in UTF-8 in which cases of plagiarism are found, plus XML files comprising meta data about them. The documents are divided into so-called source documents and suspicious documents, where suspicious documents are supposed to be analyzed for evidence of plagiarism. Each suspicious document is a priori paired with one or more source document, i.e., the task in text alignment is to extract passages of plagiarized text from a given pair of documents, if there are any. Text alignment does not involve retrieval of source documents, a different problem studied in the accompanying source retrieval task [20]. The meta data for each pair of documents reveals if and where plagiarism cases are found within in the form of character offsets and lengths for each plagiarism case. These ground truth annotations are used to measures the performance of a text alignment algorithm in extracting these plagiarism cases. While the problem is trivial for situations where text has been lifted verbatim from a source document into a suspicious document, the problem gets a lot more difficult in case the plagiarized text passage has been obfuscated, e.g., by being paraphrased, translated, or summarized. There are many ways to obfuscate a plagiarized test passage, both in the real world as well as using (semi-)automatic emulations of the real thing. Therefore, each dataset is supposed to contain an additional folder for each obfuscation strategy applied; the XML meta data files revealing the ground truth annotations are divided accordingly into these folders. To assist participants in getting the format of their datasets right, we supplied them with a format validation tool that checks formatting details and that performs basic sanity checks. This tool, of course, cannot ascertain whether the text passages annotated as plagiarized are actually meaningful or not.

4.2 Dataset Overview

Table 1 compiles an overview of the submitted datasets. The table shows the sizes of each corpus in terms of documents and plagiarism cases within. The sizes vary greatly from around 160 documents and about the same number of plagiarism cases to more than 27000 documents and more than 11000 plagiarism cases. Most of the datasets comprise English documents, whereas two feature cross-language plagiarism from Urdu and Persian to English. Two datasets contain only non-English documents in Persian and Chinese. Almost all datasets also contain a portion of suspicious documents that do not contain any plagiarism, whereas the datasets of Alvi et al. [4], Mohtaj et al. [32], and Palkovskii and Belov [36] contain only few such documents, and that of Kong et al. [27] none. The documents are mostly short (up to 10 pages), where a page is measured as 1500 chars (a norm page in print publishing), which corresponds to about 288 words in

Table 1. Overview of dataset statistics for the eight submitted datasets

Dataset Statistics	Alvi [4]	Asghari [5]	Cheema [9]	Hanif [21]	Khoshnavataher [25]	Kong [27]	Mohtaj [32]	Palkovskii [36]
Generic								
documents	272	27115	1000	1000	2111	160	4261	5057
plagiarism cases	150	11200	250	270	823	152	2781	4185
languages	en	en-fa	en	en-ur	fa	zh	en	en
Document purpose								
source documents	37%	74%	50%	50%	50%	95%	78%	64%
suspicious documents								
- with plagiarism	55%	13%	25%	27%	25%	5%	15%	33%
- w/o plagiarism	8%	13%	25%	23%	25%	0%	5%	3%
Document length								
short (<10 pp.)	92%	68%	100%	95%	95%	42%	46%	93%
medium (10-100 pp.)	8%	32%	0%	5%	5%	57%	54%	7%
long (>100 pp.)	0%	0%	0%	0%	0%	0%	0%	0%
Plagiarism per document								
hardly (<20%)	48%	68%	97%	89%	27%	75%	68%	80%
medium (20%-50%)	19%	32%	3%	10%	72%	25%	32%	19%
much (50%-80%)	33%	0%	0%	1%	0%	0%	0%	1%
entirely (>80%)	0%	0%	0%	0%	0%	0%	0%	0%
Case length								
short (<1k characters)	99%	99%	90%	99%	41%	85%	100%	97%
medium (1k-3k characters)	1%	1%	10%	1%	59%	11%	0%	3%
long (>3k characters)	0%	0%	0%	0%	0%	4%	0%	0%
Obfuscation synthesis approaches								
character-substitution	33%	–	–	–	–	–	–	–
human-retelling	33%	–	–	–	–	–	–	–
synonym-replacement	33%	–	–	–	–	–	–	–
translation-obfuscation	–	100%	–	–	–	–	–	24%
no-obfuscation	–	–	–	–	31%	–	10%	24%
random-obfuscation	–	–	–	–	69%	–	87%	24%
simulated-obfuscation	–	–	–	–	–	–	3%	–
summary-obfuscation	–	–	–	–	–	–	–	28%
undergrad	–	–	61%	–	–	–	–	–
phd	–	–	16%	–	–	–	–	–
masters	–	–	6%	–	–	–	–	–
undergrad-in-progress	–	–	17%	–	–	–	–	–
plagiarism	–	–	–	100%	–	–	–	–
real-plagiarism	–	–	–	–	–	100%	–	–

English. Some datasets also contain medium-length documents of about 10-100 pages, however, only the datasets of Kong et al. [27] and Mohtaj et al. [32] have more medium documents than short ones. No datasets have long documents; for comparison, the PAN plagiarism corpora 2009-2012 contain about 15% long documents.

The portion of plagiarism per document is below 50% of a given document in almost all cases, suggesting that the plagiarism cases are mostly short, too. This is corroborated when looking at the distributions of case length; almost all cases unanimously below 1000 characters, except for the dataset of Khoshnavataher et al. [25] and, to a lesser extent, that of Kong et al. [27] and Cheema et al. [9]. Only the dataset of Alvi et al. [4] contains documents with much (up to 80%) plagiarism. Again, no dataset contains documents that are entirely plagiarized, and only Kong et al. [27] has a small percentage of long plagiarism cases.

With regard to the obfuscation synthesis approaches, we report the names used by the dataset authors for consistency with the respective datasets' folder names. The approaches will be discussed in more detail below, however, we depart from the names used by the dataset authors for the obfuscation synthesis approaches, since they are inconsistent with the literature.

4.3 Document Sources

The first building block of every text alignment dataset is the set of documents used, which is divided into suspicious documents and source documents. One of the main obstacles in this connection is to pair suspicious documents with source documents that are roughly about the same topic, or that partially share a topic. Ideally, one would choose a set of documents that naturally possess such a relation, however, such documents are not readily available at scale. Although it is not a strict necessity to ensure topical relations for pairs of suspicious and source documents, doing so adds to the realism of an evaluation dataset for text alignment, since, in the real world, spurious similarities between topically related documents that are not plagiarism are much more likely than otherwise.

For the PAN plagiarism corpora 2009-2012, we employed documents obtained from the Project Gutenberg for the most part [50], pairing documents at random and disregarding their topical relation. We have experimented with clustering algorithms to select document pairs with at least a basic topical relation, but this had only limited success. For PAN 2013, we switched from Project Gutenberg documents to using ClueWeb09 documents, based on the Webis text reuse corpus 2012, Webis-TRC-12 [49]. In this corpus, as well as the text alignment corpus that we derived from it, a strong topical relation between documents can be assumed, since the source documents have been manually retrieved for a predefined TREC topic.

Regarding the submitted datasets, those of Asghari et al. [5], Cheema et al. [9], Hanif et al. [21], Khoshnavataher et al. [25], and Mohtaj et al. [32] employ documents drawn from Wikipedia. Cheema et al. [9] also employ documents from Project Gutenberg, but it is unclear whether pairs of suspicious and source documents are selected from both, or whether they are always from the same document source. In all cases, the authors make an effort to pair documents that are topically related. In this regard, Khoshnavataher et al. [25] and Mohtaj et al. [32] both employ the same strategy of clustering Wikipedia articles using the bipartite document-category graph and the graph-based clustering algorithm of Rosvall and Bergstrom [55]. Asghari et al. [5] rely on cross-language links between Wikipedia articles in different languages to identify documents about the same topic.

Alvi et al. [4] employs translations of Grimm's fairy tales into English, obtained from Project Gutenberg, pairing documents which have been translated by different authors. Therefore this dataset is also limited to a very specific genre comprising sometimes rather old forms of usage and style. Nevertheless, pairs of documents that tell the same fairy tale are bound to have strong topical relation.

Kong et al. [27] follow our strategy to construct the Webis-TRC-12 [49], namely they asked 10 volunteers to genuinely write as many essays on predefined topics. The volunteers were asked to use a web search engine to manually retrieve topic-related sources and to reuse text passages found on the web pages to compose their essays. This way, the resulting suspicious documents and their corresponding source documents also possess a strong topical relation.

Palkovskii and Belov [36] took a shortcut by simply reusing the training dataset of the PAN 2013 shared task on text alignment [45], simply applying an additional means of obfuscation across the corpus as detailed below.

4.4 Obfuscation Synthesis

The second building block of every text alignment dataset is the set of obfuscation approaches used to emulate human plagiarists who try to hide their plagiarism. Obfuscation of plagiarized text passages is what makes the task of text alignment difficult for detection algorithms as well as for constructing datasets. The difficulty for the latter arises from the fact that real plagiarism is hard to find at scale, especially when plagiarists invest a lot of effort in hiding it. It can be assumed that there is a bias in all plagiarism cases that make the news toward cases that are easier to be detected. Therefore, approaches have to be devised that will yield obfuscated plagiarism that comes close to the real thing, but can be created at scale with reasonable effort.

There are basically two alternatives to obfuscation synthesis, namely within context and without: within context, the entire document surrounding a plagiarized passage is created simultaneously, either manually or automatically, whereas without context, a plagiarized passage is created independently and afterwards embedded into a host document. The latter is easier to be accomplished, but lacks realism since plagiarized passages are not interleaved with the surrounding host document or other plagiarized passages around it. Moreover, when embedding an independently created plagiarized passages into a host document, the selected host document should be topically related, or else a basic topic drift analysis will reveal a plagiarized passage.

For the PAN plagiarism corpora 2009-2013, we devised a number of obfuscation approaches ranging from automatic obfuscation to manual obfuscation. This was done without context, embedding plagiarized passages into host documents after obfuscation. In particular, the obfuscation approaches are the following:

- *Random text operations.* Random shuffling, insertion, replacement, or removal of phrases and sentences. Insertions and replacements are obtained from context documents [51].
- *Semantic word variation.* Random replacement words with synonyms, antonyms, hyponyms, or hypernyms [51].
- *Part-of-speech-preserving word shuffling.* Shuffling of phrases while maintaining the original POS sequence [51].
- *Machine translation.* Automatic translation from one language to another [51].
- *Machine translation and manual copyediting.* Manually corrected output of machine translation [43].
- *Manual translation from parallel corpus.* Usage of translated passages from an existing parallel corpus [44].
- *Manual paraphrasing via crowdsourcing.* Asking human volunteers to paraphrase a given passage of text, possibly on crowdsourcing platforms, such as Amazon's Mechanical Turk [6, 41, 50].
- *Cyclic translation.* Automatic translation of a text passage from one language via a sequence of other languages to the original language [45]
- *Summarization.* Summaries of long text passages or complete documents obtained from the corpora of summaries, such as the DUC corpora [45].

Regarding the submitted datasets, Kong et al. [27] recreate our previous work on generating completely manual text reuse cases for the Webis-TRC-12 [49] on a small

scale. They asked volunteers to write essays on predefined topics, reusing text passages from the web pages they found during manual retrieval. From all submitted datasets, Kong et al. [27] present the only one where obfuscation has been synthesized within context. They also introduce an interesting twist: to maximize the obfuscation of the plagiarized text passages, the student essays have been submitted to a plagiarism detection system widely used at Chinese universities, and the volunteers have paraphrased their essays until the system could not detect the plagiarism, anymore.

For all other datasets, obfuscated plagiarized passages have been synthesized without context, and then embedded into host documents. Here, Alvi et al. [4] employ manual translations from a pseudo-parallel corpus, namely different editions of translations of Grimm's fairy tales. These translation pairs are then embedded into other, unrelated fairy tales, assuming that the genre of fairy tales in general will, to some extent, provide context with a more or less matching topic. It remains to be seen whether a topic drift analysis may reveal the plagiarized passages. At any rate, independent translations of fairy tales will provide for an interesting challenge for text alignment algorithms. In addition to that, Alvi et al. [4] also employ the above obfuscation approach of semantic word variation, and a new approach which we call UTF character substitution. Here, characters are replaced by look-alike characters from the UTF table, which makes it more difficult, though not impossible, for text alignment algorithms to match words at a lexical level. Note in this connection that Palkovskii and Belov [36] have also applied UTF character substitution on top of the reused PAN 2013 training dataset; they have already pointed out back at PAN 2009 that students sometimes apply this approach in practice [37].

Cheema et al. [9] employ manual paraphrasing via crowdsourcing; they have recruited colleagues, friends, and students at different stages of their education, namely undergrads, bachelors, masters, and PhDs, and asked them to paraphrase a total of 250 text passages selected from their respective study domain (e.g., technology, life sciences, and humanities). These paraphrased text passages have then been embedded into documents drawn from Wikipedia and Project Gutenberg which were selected according to topical similarity to the paraphrased text passages.

Hanif et al. [21] employ machine translation with and without manual copyediting, and machine translation with random text operations to obfuscate text passages obtained from the Urdu Wikipedia. The translated passage are then embedded into host documents selected so that they match the topic of the translated passages.

Since the datasets of Asghari et al. [5], Mohtaj et al. [32], and Khoshnavataher et al. [25] have been compiled by more or less the same people, their construction process if very similar. In all cases, obfuscated text passages obtained from Wikipedia articles are embedded into other Wikipedia articles that serve as suspicious documents. For the monolingual datasets Mohtaj et al. [32] and Khoshnavataher et al. [25] employ random text operations as obfuscation approach. In addition, for both monolingual and cross-language datasets, a new way of creating obfuscation is devised: Asghari et al. [5] and Mohtaj et al. [32] employ what we call "sentence stitching" with sentence pairs obtained from parallel corpora when creating cross-language plagiarism, or paraphrase corpora for monolingual plagiarism. To create a plagiarized passage and its corresponding source passage, sentence pairs from such corpora are selected and then simply ap-

pended to each other to form aligned passages of text. Various degrees of obfuscation difficulty can be introduced by measuring the similarity of sentence pairs with an appropriate similarity measure, and by combining sentences with high similarity to create low obfuscation, and vice versa. The authors try to ensure that combined sentences pairs have at least some similarity to other sentences found in a generated pair of passages by first clustering the sentences in the corpora used by their similarity. However, the success of the latter depends on how many semantically related sentence pairs are actually found in the corpora used, since clustering algorithms will even find clusters of sentences when there are only unrelated sentence pairs.

In summary, the authors of the submitted dataset propose the following new obfuscation synthesis approaches:

- *UTF character substitution.* Replacement of characters with look-alike UTF characters [4, 36].
- *Sentence stitching using parallel corpora or paraphrase corpora.* Generation of pairs of text passages from a selection of translated or paraphrased passages [5, 32].
- *Manual paraphrasing against plagiarism detection.* Paraphrasing a text passage until a plagiarism detector fails to detect the text passage as plagiarism [27].

Discussion In the first years of PAN and the plagiarism detection task, we have harvested a lot of the low-hanging fruit in terms of constructing evaluation resources, and in particular in devising obfuscation synthesis approaches. In this regard, it is not surprising that, despite the fact that eight datasets haven been submitted, only three completely new approaches have been proposed. If things were different, and we would start a task from scratch in this way, participants who decide to construct datasets would certainly have come up with most of these approaches themselves. Perhaps, by having proposed so many of the existing obfuscation synthesis approaches ourselves, we may be stifling creativity by anchoring the thoughts of participants to what is already there instead of what is missing. For example, it is interesting to note that none of the participants have actually implemented and employed automatic paraphrasing algorithms or any other form of text generation, e.g., based on language models.

5 Dataset Validation and Evaluation

Our approach at validating data submissions for shared tasks is twofold: (1) all participants who submit a dataset have been asked to peer-review the datasets of all other participants, and (2) running all 31 pieces of software that have been submitted to previous editions of our shared task on text alignment against the submitted datasets. In what follows, we review the reports of the participants' peer-reviews that have been submitted as part of their notebook papers to this year's shared task, we introduce the performance measures used to evaluate text alignment software, and we report on the evaluation results obtained from running the text alignment software against the submitted datasets using the TIRA experimentation platform.

5.1 Dataset Peer-Review

Peer-review is one of the traditional means of the scientific community to check and ensure quality. Data submissions introduce new obstacles to the successful organization of a peer-review for the following reasons:

- *Dataset size.* Datasets for shared tasks tend to be huge, which renders individual reviewers incapable of reviewing them all. Here, the selection of statistically representative subset may alleviate the problem, allowing for an estimation of the total amount of errors or other quality issues in a given dataset.
- *Assessment difficulty.* Even if the ground truth of a dataset is revealed, it may not be enough to easily understand and follow up on their construction principles of a dataset. Additional tools may be required to review problem instances at scale; in some cases, these tools need to solve the task's underlying problem themselves, e.g., to properly visualize problem instances, whereas, without visualization, the review time per problem instance may prohibitively long.
- *Reviewer bias.* Given a certain assessment difficulty for problem instances, even if the ground truth is revealed, reviewers may be biased to favor easy decisions over difficult ones.
- *Curse of variety.* While shared tasks typically tackle very clear-cut problems, the number of application domains where the task in question occurs may be huge. In these situations, it is unlikely that the reviewers available possess all the required knowledge, abilities, and experience to review and judge a given dataset with confidence.
- *Lack of motivation.* While it is fun and motivating to create a new evaluation resource, that is less so when reviewing those of others. Reviewers in shared task that invite data submissions may therefore feel less inclined to invest their time into reviewing other participants' contributions.
- *Privacy concerns.* Some reviewers may feel uncomfortable when passing open judgment onto their peers' work for fear of repercussions, especially when they find datasets to be sub-standard. However, an open discussion of the quality of evaluation resources of all kinds is an import prerequisite for progress.

All of the above obstacles have been observed in our case: some submitted datasets that are huge, comprising thousands of generated plagiarism cases; reviewing pairs of entire text documents up to dozens of pages long, and comparing plagiarism cases that may be extremely obfuscated is a laborious task, especially when no tools are around to help; some submitted datasets have been constructed in languages that none of the reviewers speak, except for those who constructed the dataset; and some of the invited reviewers apparently lacked the motivation to actually conduct a review in a useful manner.

The most comprehensive review has been submitted by volunteer reviewers, who did not submit a dataset of their own: Franco-Salvador et al. [12] systematically analyzed the submitted datasets both by computing dataset statistics and by manual inspection. The dataset statistics computed are mostly consistent with those we show in Table 1. Since most of the datasets have been submitted without further explanation by their authors, Franco-Salvador et al. [12] suggest to ask participants for short descriptions of how the datasets have been constructed in the future. Altogether, the reviewers reverse-engineer the datasets in their review, making educated guesses at how they

have been constructed and what are their document sources. Regarding the datasets of Asghari et al. [5], Cheema et al. [9], Mohtaj et al. [32], and Palkovskii and Belov [36], the reviewers find unusual synonym replacements as well as garbled text, which is probably due to the automatic obfuscation synthesis approaches used. Here, the automatic construction of datasets has its limits. Regarding datasets that are partially or completely non-English, the reviewers went to great lengths to study them, despite not being proficient in the datasets' languages: the reviewers translated non-English plagiarized passages to English using Google Translate in order to get an idea of whether paired text passages actually match with regard to their topic. This approach to overcoming the language barrier is highly commendable, and shows that improvisation can greatly enhance a reviewer's abilities. Even if the finer details of the obfuscation synthesis strategies applied in the non-English datasets are lost or skewed using Google Translate, the reviewers at least get an impression of the plagiarism cases. Altogether, the reviewers did not identify any extreme errors that invalidate any of the datasets for use in an evaluation.

The review submitted by Alvi et al. [4] has been conducted similarly to that of Franco-Salvador et al. [12]. The reviewers note inconsistencies of the annotations where character offsets and lengths do not appear to match the plagiarism cases in the datasets of Hanif et al. [21] and Mohtaj et al. [32]. Moreover, the reviewers also employ Google Translate to double-check the cross-language and non-English plagiarism cases for topical similarity. Altogether, the reviewers sometimes have difficulties in discerning the meaning of certain obfuscation synthesis names used by dataset authors, which is due to the fact that no explanation about them has been provided by the dataset authors. Again, they did not identify any detrimental errors. Furthermore, to help other reviewers in their task, Alvi et al. [4] shared a visual tool to help review plagiarism cases in the datasets.

The author lists of the datasets of Asghari et al. [5], Mohtaj et al. [32], and Khoshnavataher et al. [25] overlap, so that they decided to submit a joint review, written by Zarrabi et al. [67]. The reviewers compile some dataset statistics and report on manually reviewing 20 plagiarism cases per dataset. Despite some remarks on small errors identified, the authors do not find any systematic errors.

Finally, Cheema et al. [9] provide only a very short and superficial review, Kong et al. [27] compile only some corpus statistics without any remarks, and Palkovskii and Belov [36] did not take part in the review phase.

Discussion The outcome of the review phase of our shared task is a mixed bag. While some reviewers made an honest attempt to conduct thorough reviews, most did so only superficially. From what we learned, the datasets can be used for evaluation with some confidence, they are not systematically compromised. With hindsight, data submissions should still involve a review phase, however, there should be more time for peer-review than only one or two weeks. Also, the authors of submitted datasets should have a chance of seeing their reviews before the final submission deadline, so that they have a chance of improving their datasets. Reviewers should also be allowed to provide anonymous feedback. Nevertheless, the reviews should be published to allow later users of the datasets to get an impartial idea of its quality. Finally, once the datasets are actually used for evaluation purposes either in another shared task, or by independent researchers, the

researchers using them have a much higher motivation to actually look deep into the datasets they are using.

5.2 Plagiarism Detection Performance Measures

To assess the performance of the submitted datasets, we employ the text alignment software that has been submitted in previous years to the shared task of text alignment at PAN [50]: precision, recall, and granularity, which are combined into the plagdet score. Moreover, as of last year, we also compute case-level and document-level precision, recall, and F_1. In what follows, we recap these performance measures.

Character level performance measures Let S denote the set of plagiarism cases in the corpus, and let R denote the set of detections reported by a plagiarism detector for the suspicious documents. A plagiarism case $s = \langle s_{\text{plg}}, d_{\text{plg}}, s_{\text{src}}, d_{\text{src}} \rangle$, $s \in S$, is represented as a set $\mathbf{s}$ of references to the characters of d_{plg} and d_{src}, specifying the passages s_{plg} and s_{src}. Likewise, a plagiarism detection $r \in R$ is represented as $\mathbf{r}$. We say that r detects s iff $\mathbf{s} \sqcap \mathbf{r} \neq \emptyset$ and s_{plg} overlaps with r_{plg} and s_{src} overlaps with r_{src}. Based on this notation, precision and recall of R under S can be measured as follows:

$$prec(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup_{s \in S}(\mathbf{s} \sqcap \mathbf{r})|}{|\mathbf{r}|}, \qquad rec(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup_{r \in R}(\mathbf{s} \sqcap \mathbf{r})|}{|\mathbf{s}|},$$

$$\text{where} \quad \mathbf{s} \sqcap \mathbf{r} = \begin{cases} \mathbf{s} \cap \mathbf{r} & \text{if } r \text{ detects } s, \\ \emptyset & \text{otherwise.} \end{cases}$$

Observe that neither precision nor recall account for the fact that plagiarism detectors sometimes report overlapping or multiple detections for a single plagiarism case. This is undesirable, and to address this deficit also a detector's granularity is quantified as follows:

$$gran(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|,$$

where $S_R \subseteq S$ are cases detected by detections in R, and $R_s \subseteq R$ are detections of s; i.e., $S_R = \{s \mid s \in S \text{ and } \exists r \in R : r \text{ detects } s\}$ and $R_s = \{r \mid r \in R \text{ and } r \text{ detects } s\}$. Note further that the above three measures alone do not allow for a unique ranking among detection approaches. Therefore, the measures are combined into a single overall score as follows:

$$plagdet(S, R) = \frac{F_1}{\log_2(1 + gran(S, R))},$$

where F_1 is the equally weighted harmonic mean of precision and recall.

Case level performance measures Let S and R be defined as above. Further, let

$$S' = \{s \mid s \in S \text{ and } rec_{\text{char}}(s, R) > \tau_1 \text{ and } \exists r \in R\text{: } r \text{ detects } s \text{ and } prec_{\text{char}}(S, r) > \tau_2\}$$

denote the subset of all plagiarism cases S which have been detected with more than a threshold τ_1 in terms of character recall rec_{char} and more than a threshold τ_2 in terms of character precision $prec_{\text{char}}$. Likewise, let

$$R' = \{r \mid r \in R \text{ and } prec_{\text{char}}(S, r) > \tau_2 \text{ and } \exists s \in S\text{: } r \text{ detects } s \text{ and } rec_{\text{char}}(s, R) > \tau_1\}$$

denote the subset of all detections R which contribute to detecting plagiarism cases with more than a threshold τ_1 in terms of character recall rec_{char} and more than a threshold τ_2 in terms of character precision $prec_{\text{char}}$. Here, character recall and precision derive from the character level performance measures defined above:

$$prec_{\text{char}}(S, r) = \frac{|\bigcup_{s \in S}(\mathbf{s} \sqcap \mathbf{r})|}{|\mathbf{r}|}, \qquad rec_{\text{char}}(s, R) = \frac{|\bigcup_{r \in R}(\mathbf{s} \sqcap \mathbf{r})|}{|\mathbf{s}|}.$$

Based on this notation, we compute case level precision and recall as follows:

$$prec_{\text{case}}(S, R) = \frac{|R'|}{|R|}, \qquad rec_{\text{case}}(S, R) = \frac{|S'|}{|S|}.$$

The thresholds τ_1 and τ_2 can be used to adjust the minimal detection accuracy with regard to passage boundaries. Threshold τ_1 adjusts how accurate a plagiarism case has to be detected, whereas threshold τ_2 adjusts how accurate a plagiarism detection has to be. Beyond the minimal detection accuracy imposed by these thresholds, however, a higher detection accuracy does not contribute to case level precision and recall. If $\tau_1 \to 1$ and $\tau_2 \to 1$, the minimal required detection accuracy approaches perfection, whereas if $\tau_1 \to 0$ and $\tau_2 \to 0$, it is sufficient to report an entire document as plagiarized to achieve perfect case level precision and recall. In between these extremes, it is an open question which threshold settings are valid with regard to capturing the minimally required detection quality beyond which most users of a plagiarism detection system will not perceive improvements, anymore. Hence, we choose $\tau_1 = \tau_2 = 0.5$ as a reasonable trade off, for the time being: for case level precision, a plagiarism detection r counts a true positive detection if it contributes to detecting at least $\tau_1 = 0.5 \sim 50\%$ of a plagiarism case s, and, if at least $\tau_2 = 0.5 \sim 50\%$ of r contributes to detecting plagiarism cases. Likewise, for case level recall, a plagiarism case s counts as detected if at least 50% of s are detected, and, if a plagiarism detection r contributes to detecting s while at least 50% of r contributes to detecting plagiarism cases in general.

Document level performance measures Let S, R, and R' be defined as above. Further, let D_{plg} be the set of suspicious documents and D_{src} be the set potential source documents. Then $D_{\text{pairs}} = D_{\text{plg}} \times D_{\text{src}}$ denotes the set of possible pairs of documents that a plagiarism detector may analyze, whereas

$$D_{\text{pairs}|S} = \{(d_{\text{plg}}, d_{\text{src}}) \mid (d_{\text{plg}}, d_{\text{src}}) \in D_{\text{pairs}} \text{ and } \exists s \in S : d_{\text{plg}} \in s \text{ and } d_{\text{src}} \in s\}$$

denotes the subset of D_{pairs} whose document pairs contain the plagiarism cases S, and

$$D_{\text{pairs}|R} = \{(d_{\text{plg}}, d_{\text{src}}) \mid (d_{\text{plg}}, d_{\text{src}}) \in D_{\text{pairs}} \text{ and } \exists r \in R : d_{\text{plg}} \in r \text{ and } d_{\text{src}} \in r\}$$

denotes the corresponding subset of D_{pairs} for which plagiarism was detected in R. Likewise, $D_{\text{pairs}|R'}$ denotes the subset of D_{pairs} for which plagiarism was detected when requiring a minimal detection accuracy as per R' defined above. Based on this notation, we compute document level precision and recall as follows:

$$prec_{\text{doc}}(S, R) = \frac{|D_{\text{pairs}|S} \cap D_{\text{pairs}|R'}|}{|D_{\text{pairs}|R}|}, \qquad rec_{\text{doc}}(S, R) = \frac{|D_{\text{pairs}|S} \cap D_{\text{pairs}|R'}|}{|D_{\text{pairs}|S}|}.$$

Again, the thresholds τ_1 and τ_2 allow for adjusting the minimal required detection accuracy for R', but for document level recall, it is sufficient that at least one plagiarism case is detected beyond that accuracy in order for the corresponding document pair $(d_{\text{plg}}, d_{\text{src}})$ to be counted as true positive detection. If none of the plagiarism cases present in $(d_{\text{plg}}, d_{\text{src}})$ is detected beyond the minimal detection accuracy, it is counted as false negative, whereas if detections are made for a pair of documents in which no plagiarism case is present, it is counted as false positive.

Discussion Compared to the character level measures, the case level measures relax the fine-grained measurement of plagiarism detection quality to allow for judging a detection algorithm by its capability of "spotting" plagiarism cases reasonably well with respect to the minimum detection accuracy fixed by the thresholds τ_1 and τ_2. For example, a user who is interested in maximizing case level performance may put emphasis on the coverage of all plagiarism cases rather than the precise extraction of each individual plagiarized pair of passages. The document level measures further relax the requirements to allow for judging a detection algorithm by its capability "to raise a flag" for a given pair of documents, disregarding whether it finds all plagiarism cases contained. For example, a user who is interested in maximizing these measures puts emphasis on being made suspicious, which might lead to further, more detailed investigations. In this regard the three levels of performance measurement complement each other. To rank plagiarism detection with regard to their case level performance and their document level performance, we currently use the F_α-Measure. While the best setting of α is also still unclear, we resort to $\alpha = 1$.

5.3 Evaluation Results per Dataset

This section reports on the detection performances of 31 text alignment approaches that have been submitted to the corresponding shared task at PAN 2012-2015, when run against the eight datasets submitted to this year's PAN shared task on text alignment dataset construction. To cut a long story short, we distinguish three kinds of datasets among the submitted ones: (1) datasets that yield typical detection performance results with state-of-the-art text alignment approaches, (2) datasets that yield poor detection performance results because state-of-the-art text alignment approaches are not prepared for them, and (3) datasets that are entirely solved by at least one of the state-of-the-art text alignment approaches.

Datasets with typical results The datasets submitted by Alvi et al. [4], Cheema et al. [9], and Mohtaj et al. [32] yield typical detection performances among state-of-the-art text alignment approaches; Tables 2, 3, and 4 show the results. In all cases, the top plagdet detection performance is around 0.8, whereas F_1 at case level is around 085-0.88, and F_1 at document level at 0.86-0.9. However, the top-performing text alignment approach differs: the approach of Glinos [16] performs best on the dataset of Alvi et al. [4], whereas the approach of Oberreuter and Eiselt [35] performs best on the datasets of Cheema et al. [9] and Mohtaj et al. [32]. The latter approach ranks among the top text alignment approaches on all three datasets, including its preceding version from 2012 [34].

Table 2. Cross-year evaluation of text alignment software submissions from 2012 to 2015 with respect to performance measures at character level, case level, and document level for the submitted dataset from Alvi et al. [4].

Software Submission		**Character Level**				**Case Level**			**Document Level**		
Team	Year	$plagdet$	$prec$	rec	$gran$	$prec$	rec	F_1	$prec$	rec	F_1
Glinos [16]	2014	0.80	0.85	0.76	1.01	0.84	0.87	0.85	0.86	0.87	0.87
Oberreuter [35]	2014	0.76	0.87	0.67	1.00	0.90	0.69	0.78	0.91	0.69	0.79
Oberreuter [34]	2012	0.74	0.90	0.63	1.00	0.89	0.67	0.76	0.91	0.67	0.77
Palkovskii [40]	2014	0.67	0.67	0.66	1.00	0.69	0.75	0.72	0.76	0.75	0.75
Sanchez-Perez [56]	2015	0.66	0.86	0.53	1.00	0.78	0.53	0.63	0.78	0.53	0.63
Sanchez-Perez [57]	2014	0.62	0.90	0.48	1.00	0.85	0.49	0.62	0.85	0.49	0.62
Kong [26]	2014	0.57	0.86	0.44	1.00	0.86	0.48	0.62	0.86	0.48	0.62
Kong [28]	2013	0.57	0.80	0.46	1.00	0.88	0.49	0.63	0.88	0.49	0.63
Palkovskii [38]	2012	0.56	0.85	0.43	1.00	0.80	0.48	0.60	0.80	0.48	0.60
Gross [19]	2014	0.56	0.98	0.40	1.00	0.92	0.47	0.62	0.92	0.47	0.62
Suchomel [66]	2013	0.45	0.46	0.44	1.00	0.33	0.49	0.40	0.33	0.49	0.40
Palkovskii [39]	2013	0.44	0.70	0.34	1.05	0.53	0.40	0.45	0.62	0.40	0.49
Suchomel [65]	2012	0.41	0.82	0.27	1.00	0.71	0.32	0.44	0.71	0.32	0.44
Nourian [33]	2013	0.39	0.91	0.25	1.00	0.75	0.24	0.36	0.75	0.24	0.36
Saremi [59]	2013	0.36	0.70	0.28	1.19	0.33	0.28	0.31	0.43	0.28	0.35
R. Torrejón [52]	2012	0.28	0.96	0.17	1.00	0.43	0.16	0.23	0.43	0.16	0.23
Kueppers [30]	2012	0.26	0.92	0.15	1.00	0.65	0.15	0.24	0.65	0.15	0.24
Alvi [3]	2014	0.24	1.00	0.13	1.00	0.93	0.17	0.28	0.93	0.17	0.28
Jayapal [24]	2013	0.10	0.47	0.09	2.0	0.00	0.00	0.00	0.00	0.00	0.00
Baseline	2015	0.03	1.00	0.02	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Abnar [2]	2014	0.03	0.94	0.02	1.00	1.00	0.03	0.05	1.00	0.03	0.05
Gillam [13]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [15]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [14]	2012	–	–	–	–	–	–	–	–	–	–
Jayapal [23]	2012	–	–	–	–	–	–	–	–	–	–
Kong [29]	2012	–	–	–	–	–	–	–	–	–	–
R. Torrejón [53]	2013	–	–	–	–	–	–	–	–	–	–
R. Torrejón [54]	2014	–	–	–	–	–	–	–	–	–	–
Sánchez-Vega [58]	2012	–	–	–	–	–	–	–	–	–	–
Shrestha [61]	2013	–	–	–	–	–	–	–	–	–	–
Shrestha [60]	2014	–	–	–	–	–	–	–	–	–	–

Table 3. Cross-year evaluation of text alignment software submissions from 2012 to 2015 with respect to performance measures at character level, case level, and document level for the submitted dataset from Cheema et al. [9].

Software Submission		**Character Level**				**Case Level**			**Document Level**		
Team	Year	$plagdet$	$prec$	rec	$gran$	$prec$	rec	F_1	$prec$	rec	F_1
Oberreuter [35]	2014	0.83	0.99	0.72	1.00	0.97	0.77	0.86	0.98	0.77	0.86
Palkovskii [40]	2014	0.75	0.96	0.61	1.00	0.96	0.69	0.80	0.96	0.69	0.80
Oberreuter [34]	2012	0.74	0.99	0.59	1.00	0.94	0.61	0.74	0.94	0.61	0.74
Glinos [16]	2014	0.72	0.92	0.59	1.00	0.92	0.62	0.74	0.95	0.62	0.75
Kong [28]	2013	0.61	0.87	0.46	1.00	0.97	0.54	0.70	0.97	0.54	0.70
Kong [26]	2014	0.59	0.93	0.43	1.00	0.85	0.47	0.61	0.85	0.47	0.61
Suchomel [66]	2013	0.59	0.96	0.42	1.00	0.90	0.50	0.64	0.90	0.50	0.64
Palkovskii [38]	2012	0.56	0.98	0.40	1.00	0.90	0.43	0.59	0.90	0.43	0.59
Suchomel [65]	2012	0.55	0.98	0.38	1.00	0.93	0.44	0.60	0.93	0.44	0.60
Palkovskii [39]	2013	0.48	1.00	0.33	1.06	0.66	0.39	0.49	0.70	0.39	0.50
Sanchez-Perez [56]	2015	0.48	0.89	0.33	1.00	0.78	0.35	0.49	0.78	0.35	0.49
Gross [19]	2014	0.45	1.00	0.30	1.02	0.84	0.34	0.48	0.86	0.34	0.49
Sanchez-Perez [57]	2014	0.45	0.89	0.30	1.00	0.81	0.33	0.47	0.81	0.33	0.47
Saremi [59]	2013	0.44	0.97	0.32	1.11	0.59	0.34	0.43	0.66	0.34	0.45
R. Torrejón [52]	2012	0.41	1.00	0.26	1.00	0.60	0.26	0.36	0.60	0.26	0.36
Alvi [3]	2014	0.32	1.00	0.20	1.05	0.56	0.18	0.28	0.61	0.18	0.28
Kueppers [30]	2012	0.30	0.88	0.18	1.00	0.56	0.18	0.27	0.56	0.18	0.28
Abnar [2]	2014	0.27	0.90	0.16	1.00	1.00	0.17	0.30	1.00	0.17	0.30
Nourian [33]	2013	0.17	0.92	0.09	1.00	0.87	0.10	0.18	0.87	0.10	0.18
Jayapal [24]	2013	0.13	0.97	0.12	2.2	0.01	0.07	0.01	0.01	0.07	0.02
Gillam [15]	2014	0.12	0.98	0.06	1.00	0.90	0.07	0.13	0.90	0.07	0.13
Baseline	2015	0.11	1.00	0.08	1.50	0.19	0.06	0.10	0.28	0.06	0.10
Gillam [14]	2012	–	–	–	–	–	–	–	–	–	–
Gillam [13]	2013	–	–	–	–	–	–	–	–	–	–
Jayapal [23]	2012	–	–	–	–	–	–	–	–	–	–
Kong [29]	2012	–	–	–	–	–	–	–	–	–	–
R. Torrejón [53]	2013	–	–	–	–	–	–	–	–	–	–
R. Torrejón [54]	2014	–	–	–	–	–	–	–	–	–	–
Sánchez-Vega [58]	2012	–	–	–	–	–	–	–	–	–	–
Shrestha [61]	2013	–	–	–	–	–	–	–	–	–	–
Shrestha [60]	2014	–	–	–	–	–	–	–	–	–	–

Table 4. Cross-year evaluation of text alignment software submissions from 2012 to 2015 with respect to performance measures at character level, case level, and document level for the submitted dataset from Mohtaj et al. [32].

Software Submission		**Character Level**				**Case Level**			**Document Level**		
Team	Year	$plagdet$	$prec$	rec	$gran$	$prec$	rec	F_1	$prec$	rec	F_1
Oberreuter [35]	2014	0.80	0.83	0.78	1.00	0.90	0.85	0.88	0.95	0.85	0.90
Oberreuter [34]	2012	0.76	0.79	0.74	1.00	0.82	0.80	0.81	0.85	0.82	0.83
Sanchez-Perez [57]	2014	0.76	0.81	0.71	1.00	0.95	0.78	0.86	0.97	0.82	0.89
Sanchez-Perez [56]	2015	0.75	0.80	0.71	1.00	0.94	0.78	0.86	0.97	0.84	0.90
Gross [19]	2014	0.73	0.91	0.61	1.01	0.96	0.73	0.83	0.98	0.78	0.87
Palkovskii [40]	2014	0.73	0.86	0.62	1.00	0.95	0.70	0.81	0.97	0.78	0.86
Kong [26]	2014	0.66	0.76	0.59	1.00	0.86	0.68	0.76	0.92	0.75	0.83
Kong [28]	2013	0.66	0.76	0.59	1.00	0.86	0.68	0.76	0.92	0.75	0.83
R. Torrejón [52]	2012	0.65	0.59	0.74	1.00	0.49	0.80	0.61	0.49	0.77	0.60
Suchomel [66]	2013	0.65	0.73	0.59	1.00	0.66	0.67	0.66	0.70	0.70	0.70
Kueppers [30]	2012	0.63	0.82	0.51	1.00	0.85	0.60	0.71	0.91	0.71	0.80
Saremi [59]	2013	0.61	0.86	0.60	1.24	0.62	0.70	0.66	0.87	0.76	0.81
Palkovskii [38]	2012	0.59	0.76	0.48	1.00	0.77	0.54	0.63	0.81	0.62	0.71
Suchomel [65]	2012	0.59	0.82	0.46	1.00	0.72	0.52	0.60	0.76	0.56	0.65
Abnar [2]	2014	0.56	0.92	0.42	1.01	0.92	0.44	0.60	0.97	0.56	0.72
Palkovskii [39]	2013	0.56	0.88	0.43	1.04	0.62	0.47	0.53	0.78	0.57	0.67
Glinos [16]	2014	0.54	0.95	0.38	1.01	0.97	0.39	0.55	0.99	0.47	0.64
Alvi [3]	2014	0.46	0.94	0.31	1.03	0.81	0.36	0.50	0.89	0.48	0.62
Nourian [33]	2013	0.36	0.88	0.23	1.00	0.90	0.25	0.39	0.91	0.32	0.47
Jayapal [24]	2013	0.22	0.86	0.23	2.14	0.03	0.12	0.05	0.11	0.15	0.12
Baseline	2015	0.18	0.93	0.12	1.24	0.21	0.09	0.12	0.22	0.10	0.14
Gillam [15]	2014	0.06	0.84	0.03	1.00	0.94	0.03	0.07	0.97	0.05	0.09
Gillam [13]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [14]	2012	–	–	–	–	–	–	–	–	–	–
Jayapal [23]	2012	–	–	–	–	–	–	–	–	–	–
Kong [29]	2012	–	–	–	–	–	–	–	–	–	–
R. Torrejón [53]	2013	–	–	–	–	–	–	–	–	–	–
R. Torrejón [54]	2014	–	–	–	–	–	–	–	–	–	–
Sánchez-Vega [58]	2012	–	–	–	–	–	–	–	–	–	–
Shrestha [61]	2013	–	–	–	–	–	–	–	–	–	–
Shrestha [60]	2014	–	–	–	–	–	–	–	–	–	–

For comparison, the winning text alignment approach of PAN 2014 from Sanchez-Perez et al. [57], as well as its 2015 successor [56], achieves mid-range performances on the dataset of Alvi et al. [4], low performance on that of Cheema et al. [9], and second rank, following the approaches of Oberreuter on the dataset of Mohtaj et al. [32]. Some of these performance differences may be attributed to the fact that the approach of Sanchez-Perez et al. [57] has been optimized to work well on the previous year's PAN plagiarism corpus, whereas it has not, yet, been optimized against the submitted datasets, nor against all of them in combination.

Apparently, the obfuscation synthesis approaches applied during construction of the three aforementioned datasets compare in terms of difficulty to the ones applied during construction of the PAN plagiarism corpora. The detection performances on the submitted datasets are not perfect so that further algorithmic improvements are required. These datasets complement the ones that are already used, and, since they have been constructed independently, while still allowing for the existing text alignment approaches to work well, they verify that the previous datasets which have been constructed exclusively by ourselves, are fit for purpose.

Datasets with perfect results For one of the submitted datasets, the text alignment approaches exert an odd performance characteristic, that is: either they detect almost all plagiarism, or none at all. Table 5 shows the performances obtained on the dataset of Khoshnavataher et al. [25]. The plagdet performances of the eight top-performing text alignment approaches range from 0.89 to 0.98, the top-performing one being the approach of Glinos [16]. At case level and at document level, all of them achieve F_1 scores above 0.9. There are only four approaches that achieve mid-range performances, including the Baseline, whereas the performances of all other approaches is negligible. The Baseline implements a basic text alignment approach using 4-grams for seeding, and rule-based merging.

The dataset of Khoshnavataher et al. [25] comprises Persian documents only, so that this performance characteristic demonstrates which of the approaches can cope with this language and which cannot. An explanation for the fact that not all approaches work may be that some work at the lexical level, whereas others employ sophisticated linguistic processing pipelines that may not be adapted to processing Persian text. Moreover, the fact that those approaches which are capable of detecting plagiarism within Persian documents do so almost flawlessly, hints that the obfuscation synthesis approaches applied by Khoshnavataher et al. [25] do not seem to yield notable obfuscation.

From reviewing the notebooks of the respective approaches, however, it is not entirely clear whether they indeed work at the lexical level. Glinos [16] mentions some pre-processing at word-level, but applies the character-based Smith-Waterman algorithm to align text passages between a pair of documents. Palkovskii and Belov [38], who provides the second-best performing approach on Khoshnavataher et al. [25]'s dataset, report to employ a basic Euclidian distance-based clustering approach in their approach, which also hints that no linguistic pre-processing is applied. Interestingly, the best-performing approach of PAN 2014 from Sanchez-Perez et al. [57] does not work well, whereas this year's refined version Sanchez-Perez et al. [56] is ranked third. This suggests that the authors added a fallback solution for situations where only lexical matching applies, e.g., based on their approach to predict on-the-fly what kind of

Table 5. Cross-year evaluation of text alignment software submissions from 2012 to 2015 with respect to performance measures at character level, case level, and document level for the submitted dataset from Khoshnavataher et al. [25].

Software Submission		**Character Level**				**Case Level**			**Document Level**		
Team	Year	$plagdet$	$prec$	rec	$gran$	$prec$	rec	F_1	$prec$	rec	F_1
Glinos [16]	2014	0.98	0.97	0.99	1.00	0.98	0.99	0.99	0.98	0.99	0.99
Palkovskii [38]	2012	0.95	0.94	0.95	1.00	0.98	0.96	0.97	0.98	0.96	0.97
Sanchez-Perez [56]	2015	0.94	0.91	0.96	1.00	0.97	0.98	0.98	0.98	0.98	0.98
Suchomel [65]	2012	0.93	0.93	0.92	1.00	0.93	0.97	0.95	0.93	0.97	0.95
Palkovskii [40]	2014	0.92	0.88	0.95	1.00	0.94	0.97	0.96	0.94	0.97	0.96
Palkovskii [39]	2013	0.91	0.89	0.92	1.00	0.89	0.95	0.92	0.96	0.95	0.96
Suchomel [66]	2013	0.91	0.86	0.95	1.00	0.88	0.99	0.93	0.88	0.99	0.93
Alvi [3]	2014	0.89	0.95	0.90	1.05	0.91	0.95	0.93	0.99	0.95	0.97
Gross [19]	2014	0.80	0.76	0.98	1.09	0.72	1.00	0.84	0.78	1.00	0.88
Abnar [2]	2014	0.55	0.73	0.45	1.01	0.57	0.46	0.52	0.63	0.46	0.53
Oberreuter [35]	2014	0.40	0.25	1.00	1.00	0.13	1.00	0.23	0.15	1.00	0.25
Baseline	2015	0.37	0.97	0.50	2.41	0.14	0.37	0.20	0.34	0.37	0.35
Kong [26]	2014	0.07	0.04	0.80	1.00	0.00	0.80	0.00	0.00	0.80	0.00
Kong [28]	2013	0.07	0.04	0.80	1.00	0.00	0.80	0.00	0.00	0.80	0.00
Oberreuter [34]	2012	0.06	0.33	0.03	1.00	0.14	0.03	0.05	0.14	0.03	0.05
Sanchez-Perez [57]	2014	0.01	0.62	0.01	1.00	0.14	0.00	0.01	0.14	0.00	0.01
Saremi [59]	2013	0.00	0.09	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
R. Torrejón [52]	2012	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Jayapal [24]	2013	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Kueppers [30]	2012	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Nourian [33]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [14]	2012	–	–	–	–	–	–	–	–	–	–
Gillam [13]	2013	–	–	–	–	–	–	–	–	–	–
Gillam [15]	2014	–	–	–	–	–	–	–	–	–	–
Jayapal [23]	2012	–	–	–	–	–	–	–	–	–	–
Kong [29]	2012	–	–	–	–	–	–	–	–	–	–
R. Torrejón [53]	2013	–	–	–	–	–	–	–	–	–	–
R. Torrejón [54]	2014	–	–	–	–	–	–	–	–	–	–
Sánchez-Vega [58]	2012	–	–	–	–	–	–	–	–	–	–
Shrestha [61]	2013	–	–	–	–	–	–	–	–	–	–
Shrestha [60]	2014	–	–	–	–	–	–	–	–	–	–

obfuscation is at hand to adjust their detection approach accordingly. Since the dataset of Khoshnavataher et al. [25] apparently does not comprise noteworthy obfuscation, which results in mostly verbatim overlap of text passages between a given pair of suspicious document and source document, this may trigger the classification approach used by Sanchez-Perez et al. [56], which then applies a basic algorithm that deals with such situations at the lexical level.

Another noteworthy issue about this dataset is that some of its reviewers note that it has a high quality, which may hint at reviewer bias toward plagiarism cases that are more easy to be detected, compared to ones that are difficult to identify, even for a human. In this regard, implementing obfuscation synthesis approaches, which are supposed to construct difficult plagiarism cases, also makes the task of reviewing their results a lot more difficult, so that reviewers may tend to favor easy decisions over the difficult ones.

Datasets with poor results On four of the submitted datasets from Kong et al. [27], Asghari et al. [5], Palkovskii and Belov [36], and Hanif et al. [21], the text alignment approaches perform poorly, detecting almost none of the plagiarism cases. The best performances are obtained by the two versions of the approach of Oberreuter and Eiselt [35] and the two versions of the approach of Suchomel et al. [66] on the dataset comprising Chinese plagiarism cases from Kong et al. [27]. However, the top plagdet score achieved is still only 0.18. Most of the text alignment approaches detect almost none of the plagiarism cases on this dataset, whereas all approaches fail on the remaining datasets from Asghari et al. [5], Palkovskii and Belov [36], and Hanif et al. [21].

This does not hint at any flaws in the datasets, but is testimony to the datasets' difficulty. The datasets of Asghari et al. [5] and Hanif et al. [21] are the only ones comprising cross-language plagiarism from Persian and Urdu to English, respectively. Since no lexical similarity between these languages can be expected, apart from perhaps a few named entities, and since apparently none of the text alignment approaches feature any kind of translation module or cross-language similarity measure, they cannot cope with these kinds of plagiarism cases.

Regarding the dataset submitted by Kong et al. [27], which comprises monolingual Chinese plagiarism cases, a set of text alignment approaches seems to work, to a small extent, that corresponds to those working on the dataset of Khoshnavataher et al. [25], probably for similar reasons as outlined above. However, for their dataset, Kong et al. [27] have optimized the obfuscation of the plagiarism cases until a Chinese plagiarism detector was unable to detect them, which makes the task of detecting these plagiarism cases very difficult, since lexical similarities are not to be expected. Moreover, most of the existing approaches are probably not optimized to process Chinese text, since each letter may carry a lot more semantics than letters from the Latin alphabet. Therefore, shorter character sequences may already hint a significant semantic similarity.

Regarding the dataset of Palkovskii and Belov [36], the obfuscation approach of substituting characters with look-alike UTF characters seems to successfully confound all of the existing text alignment approaches. The dataset of Alvi et al. [4], where the performances have otherwise been typical, also contains plagiarism cases where this kind of obfuscation has been applied.

Table 6. Cross-year evaluation of text alignment software submissions from 2012 to 2015 with respect to performance measures at character level, case level, and document level for the submitted dataset from Kong et al. [27].

Software Submission		**Character Level**				**Case Level**			**Document Level**		
Team	Year	$plagdet$	$prec$	rec	$gran$	$prec$	rec	F_1	$prec$	rec	F_1
Oberreuter [35]	2014	0.18	0.16	0.21	1.00	0.16	0.23	0.19	0.21	0.23	0.22
Suchomel [66]	2013	0.17	0.16	0.18	1.00	0.10	0.19	0.13	0.12	0.18	0.14
Suchomel [65]	2012	0.16	0.17	0.15	1.00	0.11	0.16	0.13	0.14	0.15	0.15
Oberreuter [34]	2012	0.12	0.14	0.11	1.00	0.11	0.11	0.11	0.14	0.11	0.12
Gross [19]	2014	0.11	0.13	0.10	1.08	0.13	0.11	0.12	0.23	0.12	0.16
Abnar [2]	2014	0.07	0.21	0.08	2.17	0.10	0.11	0.10	0.25	0.14	0.18
Alvi [3]	2014	0.07	0.15	0.08	2.0	0.04	0.09	0.06	0.17	0.14	0.15
Palkovskii [38]	2012	0.06	0.10	0.05	1.00	0.07	0.05	0.06	0.07	0.05	0.06
Glinos [16]	2014	0.06	0.07	0.05	1.00	0.04	0.04	0.04	0.11	0.07	0.08
Kong [26]	2014	0.05	0.03	0.56	1.01	0.00	0.57	0.01	0.01	0.56	0.03
Baseline	2015	0.05	0.18	0.06	2.68	0.00	0.05	0.01	0.02	0.08	0.03
Kong [28]	2013	0.05	0.03	0.56	1.01	0.00	0.56	0.01	0.01	0.55	0.03
Sanchez-Perez [57]	2014	0.04	0.16	0.02	1.00	0.20	0.03	0.05	0.25	0.01	0.03
R. Torrejón [52]	2012	0.03	0.04	0.02	1.00	0.06	0.03	0.04	0.06	0.04	0.05
Palkovskii [39]	2013	0.02	0.10	0.01	1.00	0.00	0.02	0.00	0.00	0.01	0.00
Saremi [59]	2013	0.02	0.04	0.02	2.62	0.00	0.02	0.00	0.04	0.03	0.03
Palkovskii [40]	2014	0.01	0.15	0.01	1.00	0.17	0.01	0.02	0.20	0.01	0.03
Sanchez-Perez [56]	2015	0.01	0.50	0.01	1.00	0.50	0.01	0.02	1.00	0.01	0.03
Jayapal [24]	2013	0.00	0.08	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [15]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Kueppers [30]	2012	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Nourian [33]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [14]	2012	–	–	–	–	–	–	–	–	–	–
Gillam [13]	2013	–	–	–	–	–	–	–	–	–	–
Jayapal [23]	2012	–	–	–	–	–	–	–	–	–	–
Kong [29]	2012	–	–	–	–	–	–	–	–	–	–
R. Torrejón [53]	2013	–	–	–	–	–	–	–	–	–	–
R. Torrejón [54]	2014	–	–	–	–	–	–	–	–	–	–
Sánchez-Vega [58]	2012	–	–	–	–	–	–	–	–	–	–
Shrestha [61]	2013	–	–	–	–	–	–	–	–	–	–
Shrestha [60]	2014	–	–	–	–	–	–	–	–	–	–

Table 7. Cross-year evaluation of text alignment software submissions from 2012 to 2015 with respect to performance measures at character level, case level, and document level for the submitted dataset from Asghari et al. [5].

Software Submission		**Character Level**				**Case Level**			**Document Level**		
Team	Year	$plagdet$	$prec$	rec	$gran$	$prec$	rec	F_1	$prec$	rec	F_1
Oberreuter [34]	2012	0.03	0.18	0.02	1.00	0.08	0.02	0.03	0.08	0.02	0.03
Sanchez-Perez [57]	2014	0.00	0.71	0.00	1.00	0.28	0.00	0.00	0.28	0.00	0.00
Kong [28]	2013	0.00	0.01	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Kong [26]	2014	0.00	0.01	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Jayapal [24]	2013	0.00	0.89	0.00	1.12	0.00	0.00	0.00	0.00	0.00	0.00
Palkovskii [38]	2012	0.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
Suchomel [65]	2012	0.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
Suchomel [66]	2013	0.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
Alvi [3]	2014	0.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
Palkovskii [40]	2014	0.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
Sanchez-Perez [56]	2015	0.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
Abnar [2]	2014	0.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
Saremi [59]	2013	0.00	0.06	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Palkovskii [39]	2013	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Baseline	2015	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Glinos [16]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Kueppers [30]	2012	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gross [19]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Nourian [33]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Oberreuter [35]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
R. Torrejón [52]	2012	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [14]	2012	–	–	–	–	–	–	–	–	–	–
Gillam [13]	2013	–	–	–	–	–	–	–	–	–	–
Gillam [15]	2014	–	–	–	–	–	–	–	–	–	–
Jayapal [23]	2012	–	–	–	–	–	–	–	–	–	–
Kong [29]	2012	–	–	–	–	–	–	–	–	–	–
R. Torrejón [53]	2013	–	–	–	–	–	–	–	–	–	–
R. Torrejón [54]	2014	–	–	–	–	–	–	–	–	–	–
Sánchez-Vega [58]	2012	–	–	–	–	–	–	–	–	–	–
Shrestha [61]	2013	–	–	–	–	–	–	–	–	–	–
Shrestha [60]	2014	–	–	–	–	–	–	–	–	–	–

Table 8. Cross-year evaluation of text alignment software submissions from 2012 to 2015 with respect to performance measures at character level, case level, and document level for the submitted dataset from Palkovskii and Belov [36].

Software Submission		**Character Level**				**Case Level**			**Document Level**		
Team	Year	$plagdet$	$prec$	rec	$gran$	$prec$	rec	F_1	$prec$	rec	F_1
Oberreuter [35]	2014	0.03	0.99	0.01	1.00	0.97	0.01	0.03	0.96	0.02	0.03
Oberreuter [34]	2012	0.03	0.96	0.01	1.00	0.90	0.01	0.03	0.93	0.02	0.03
Glinos [16]	2014	0.01	0.95	0.01	1.00	0.41	0.00	0.01	0.40	0.00	0.01
Palkovskii [40]	2014	0.01	0.87	0.01	1.07	0.62	0.01	0.01	0.71	0.01	0.01
Gross [19]	2014	0.00	1.00	0.00	1.00	0.71	0.00	0.00	0.83	0.00	0.01
Palkovskii [39]	2013	0.00	1.00	0.00	1.00	0.44	0.00	0.00	0.43	0.00	0.00
Palkovskii [38]	2012	0.00	0.97	0.00	1.00	0.67	0.00	0.00	0.67	0.00	0.00
Saremi [59]	2013	0.00	0.24	0.00	1.11	0.00	0.00	0.00	0.00	0.00	0.00
Sanchez-Perez [57]	2014	0.00	0.97	0.00	1.00	0.67	0.00	0.00	0.67	0.00	0.00
R. Torrejón [54]	2014	0.00	0.99	0.00	1.00	0.67	0.00	0.00	0.67	0.00	0.00
Suchomel [66]	2013	0.00	1.00	0.00	1.00	0.50	0.00	0.00	0.50	0.00	0.00
Kueppers [30]	2012	0.00	0.60	0.00	1.33	0.50	0.00	0.00	0.25	0.00	0.00
R. Torrejón [53]	2013	0.00	0.99	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
Suchomel [65]	2012	0.00	1.00	0.00	1.00	0.67	0.00	0.00	0.67	0.00	0.00
Alvi [3]	2014	0.00	0.83	0.00	1.25	0.17	0.00	0.00	0.25	0.00	0.00
Baseline	2015	0.00	0.89	0.00	1.33	0.11	0.00	0.00	0.20	0.00	0.00
R. Torrejón [52]	2012	0.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
Sanchez-Perez [56]	2015	0.00	0.94	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
Jayapal [24]	2013	0.00	1.00	0.00	1.25	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [13]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [15]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Kong [28]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Kong [26]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Nourian [33]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Abnar [2]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [14]	2012	–	–	–	–	–	–	–	–	–	–
Jayapal [23]	2012	–	–	–	–	–	–	–	–	–	–
Kong [29]	2012	–	–	–	–	–	–	–	–	–	–
Sánchez-Vega [58]	2012	–	–	–	–	–	–	–	–	–	–
Shrestha [61]	2013	–	–	–	–	–	–	–	–	–	–
Shrestha [60]	2014	–	–	–	–	–	–	–	–	–	–

Table 9. Cross-year evaluation of text alignment software submissions from 2012 to 2015 with respect to performance measures at character level, case level, and document level for the submitted dataset from Hanif et al. [21].

Software Submission		**Character Level**				**Case Level**			**Document Level**		
Team	Year	$plagdet$	$prec$	rec	$gran$	$prec$	rec	F_1	$prec$	rec	F_1
Saremi [59]	2013	0.02	0.28	0.02	1.60	0.00	0.02	0.00	0.00	0.02	0.00
Sanchez-Perez [57]	2014	0.02	0.30	0.01	1.00	0.00	0.01	0.00	0.00	0.01	0.00
R. Torrejón [52]	2012	0.02	1.00	0.01	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Oberreuter [35]	2014	0.01	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Jayapal [24]	2013	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Alvi [3]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Glinos [16]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Suchomel [65]	2012	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Kong [28]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Kong [26]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Kueppers [30]	2012	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gross [19]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Nourian [33]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Oberreuter [34]	2012	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Palkovskii [38]	2012	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Palkovskii [39]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Palkovskii [40]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Baseline	2015	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Sanchez-Perez [56]	2015	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Suchomel [66]	2013	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Abnar [2]	2014	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Gillam [14]	2012	–	–	–	–	–	–	–	–	–	–
Gillam [13]	2013	–	–	–	–	–	–	–	–	–	–
Gillam [15]	2014	–	–	–	–	–	–	–	–	–	–
Jayapal [23]	2012	–	–	–	–	–	–	–	–	–	–
Kong [29]	2012	–	–	–	–	–	–	–	–	–	–
R. Torrejón [53]	2013	–	–	–	–	–	–	–	–	–	–
R. Torrejón [54]	2014	–	–	–	–	–	–	–	–	–	–
Sánchez-Vega [58]	2012	–	–	–	–	–	–	–	–	–	–
Shrestha [61]	2013	–	–	–	–	–	–	–	–	–	–
Shrestha [60]	2014	–	–	–	–	–	–	–	–	–	–

Discussion The evaluation of the existing text alignment approaches on the submitted datasets leaves us with more confidence in their quality. The datasets on which the approaches perform with results comparable to their performances on the PAN plagiarism corpora mutually verify that neither dataset is too far off from the others. Since the datasets have been constructed independently, this suggest that the intuition of the authors who constructed them corresponds to ours, albeit we may have strongly influenced them with our prior work.

Regarding the datasets where the existing text alignment approaches fail, we cannot deduce that the datasets are flawed. Rather, the characteristics of the datasets suggest that either tailored detection modules are required, or rather a better abstraction of the problem domain to allow for generic approaches. Finally, regarding the dataset where a number of text alignment approaches achieve almost perfect performance, this dataset has been beaten, which means that it may be used to confirm basic capabilities of a text alignment approach, but it does inform further research.

5.4 Analysis of Execution Errors

Not all of the existing text alignment approaches work without execution errors on the submitted datasets. On some datasets, approaches fail or print error output for various reasons. Table 10 gives an overview of common errors as reported by the output of failing approaches. Most of the errors observed hint at internal software errors, whereas others are opaque because hardly any error messages are printed by the software. Despite printing error messages, some pieces of software still generate output that can be evaluated, whereas others do not. Many of the errors observed occur on the datasets containing non-English documents, but also on datasets that make use letters from non-Latin alphabets. Moreover, we have excluded a number of approaches for being too slow.

We have repeatedly tried to get the respective approaches to work, however, the errors prevailed. We have also considered to invite the original authors to fix the errors in their software, but refrained from doing so, since then the already successfully obtained performance results on other evaluation corpora would be invalidated. Changing a software after it has been submitted to a shared task, even if only a small execution error is fixed, may have side effects on the software's performance when it is re-executed on a given dataset compared to its original performance. Arguably, a submitted software should be kept a fixture for the future and participants should rather be invited to submit a new version of their software where the errors are fixed and where they are free to make other improvements as they see fit.

6 Conclusion and Outlook

In conclusion, we can say that data submissions to shared tasks are a viable option of creating evaluation resources. The datasets that have been submitted to our shared task have a variety that we would not have been able to create on our own. Despite the important role that the organizers of a shared task play in keeping things together, and

Table 10. Overview of executions errors observed on the submitted datasets.

Software Submission		**Alvi [4]**		**Cheema [9]**		**Khoshnavataher [25]**		**Mohtaj [32]**	
Team	Year		**Asghari [5]**		**Hanif [21]**		**Kong [27]**		**Palkovskii [36]**
Alvi [3]	2014	✓	internal	✓	internal	✓	✓	internal	✓
Gillam [14]	2012	runtime	runtime	runtime	runtime	runtime	runtime	runtime	runtime
Gillam [13]	2013	internal	internal	internal	internal	internal	internal	internal	internal
Gillam [15]	2014	no output	no output	no output	no output	no output	no output	no output	no output
Glinos [16]	2014	✓	internal	✓	internal	✓	✓	internal	internal
Jayapal [23]	2012	runtime	runtime	runtime	runtime	runtime	runtime	runtime	runtime
Kong [29]	2012	runtime	runtime	runtime	runtime	runtime	runtime	runtime	runtime
Kueppers [30]	2012	✓	internal	✓	internal	internal	internal	✓	✓
Oberreuter [35]	2014	✓	memory	✓	✓	✓	✓	✓	✓
R. Torrejón [52]	2012	✓	internal	✓	internal	internal	internal	✓	✓
R. Torrejón [53]	2013	internal	internal	internal	internal	internal	internal	internal	no output
R. Torrejón [54]	2014	internal	internal	internal	internal	internal	internal	internal	✓
Sánchez-Vega [58]	2012	runtime	runtime	runtime	runtime	runtime	runtime	runtime	runtime
Shrestha [61]	2013	runtime	runtime	runtime	runtime	runtime	runtime	runtime	runtime
Shrestha [60]	2014	runtime	runtime	runtime	runtime	runtime	runtime	runtime	runtime

making sure that all moving parts fit together, it is curious that data submissions are so uncommon.

In our case, we have been able to validate and evaluate the submitted datasets not only by manual peer-review, but also by executing all software submissions to previous editions of our shared task on the submitted datasets. Perhaps, asking for data submissions in a pilot task, where no software has been developed, yet, is not so attractive. It remains to be seen whether data submissions can only be successful in connection with software submissions. The organizational procedures that we have outlined in this paper may serve as first draft of a recipe to successful data submissions. However, in other contexts and other research fields, data submissions may not be as straightforward to be implemented. For example, if the data is sensitive, if it raises privacy concerns, or if it is difficult to obtain, data submissions may not be possible.

With software submissions and data submissions, we are only one step short of involving participants in all aspects of a shared tasks; what is still missing are the performance measures. Here, the organizers of a shared task often decide on a set of performance measures, whereas the community around a shared task may have different ideas as to how to measures performance. Involving participants in performance measure development and result analysis seems to be an obvious next step. Here, for example, it may well be possible to invite theoretical contributions to a shared task, since the development of performance measures for a given shared task is closely related to developing a theory around it.

Acknowledgements

We thank the participating teams of this shared task as well as those of previous editions for their devoted work.

Bibliography

1. Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, Position Papers, April 20-21, 2007. The Ohio State University,

Arlington, Virginia, USA (2007)
2. Abnar, S., Dehghani, M., Zamani, H., Shakery, A.: Expanded N-Grams for Semantic Text Alignment—Notebook for PAN at CLEF 2014. In: [7]
3. Alvi, F., Stevenson, M., Clough, P.: Hashing and Merging Heuristics for Text Reuse Detection—Notebook for PAN at CLEF 2014. In: [7]
4. Alvi, F., Stevenson, M., Clough, P.: The Short Stories Corpus—Notebook for PAN at CLEF 2015. In: [8]
5. Asghari, H., Khoshnavataher, K., Fatemi, O., Faili, H.: Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus—Notebook for PAN at CLEF 2015. In: [8]
6. Burrows, S., Potthast, M., Stein, B.: Paraphrase Acquisition via Crowdsourcing and Machine Learning. Transactions on Intelligent Systems and Technology (ACM TIST) 4(3), 43:1–43:21 (Jun 2013), http://dl.acm.org/citation.cfm?id=2483676
7. Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.): CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK. CEUR Workshop Proceedings, CEUR-WS.org (2014), http://www.clef-initiative.eu/publication/working-notes
8. Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.): CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France. CEUR Workshop Proceedings, CEUR-WS.org (2015), http://www.clef-initiative.eu/publication/working-notes
9. Cheema, W., Najib, F., Ahmed, S., Bukhari, S., Sittar, A., Nawab, R.: A Corpus for Analyzing Text Reuse by People of Different Groups—Notebook for PAN at CLEF 2015. In: [8]
10. Forner, P., Karlgren, J., Womser-Hacker, C. (eds.): CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy (2012), http://www.clef-initiative.eu/publication/working-notes
11. Forner, P., Navigli, R., Tufis, D. (eds.): CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain (2013), http://www.clef-initiative.eu/publication/working-notes
12. Franco-Salvador, M., Bensalem, I., Flores, E., Gupta, P., Rosso, P.: PAN 2015 Shared Task on Plagiarism Detection: Evaluation of Corpora for Text Alignment—Notebook for PAN at CLEF 2015. In: [8]
13. Gillam, L.: Guess Again and See if They Line Up: Surrey's Runs at Plagiarism Detection—Notebook for PAN at CLEF 2013. In: [11]
14. Gillam, L., Newbold, N., Cooke, N.: Educated Guesses and Equality Judgements: Using Search Engines and Pairwise Match for External Plagiarism Detection—Notebook for PAN at CLEF 2012. In: [10], http://www.clef-initiative.eu/publication/working-notes
15. Gillam, L., Notley, S.: Evaluating Robustness for 'IPCRESS': Surrey's Text Alignment for Plagiarism Detection—Notebook for PAN at CLEF 2014. In: [7]
16. Glinos, D.: A Hybrid Architecture for Plagiarism Detection—Notebook for PAN at CLEF 2014. In: [7]
17. Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Recent Trends in Digital Text Forensics and its Evaluation. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative (CLEF 13). pp. 282–302. Springer, Berlin Heidelberg New York (Sep 2013)
18. Gollub, T., Stein, B., Burrows, S.: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Hersh, B., Callan, J., Maarek, Y., Sanderson, M. (eds.) 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12). pp. 1125–1126. ACM (Aug 2012)

19. Gross, P., Modaresi, P.: Plagiarism Alignment Detection by Merging Context Seeds—Notebook for PAN at CLEF 2014. In: [7]
20. Hagen, M., Potthast, M., Stein, B.: Source Retrieval for Plagiarism Detection from Large Web Corpora: Recent Approaches. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2015), http://www.clef-initiative.eu/publication/working-notes
21. Hanif, I., Nawab, A., Arbab, A., Jamshed, H., Riaz, S., Munir, E.: Cross-Language Urdu-English (CLUE) Text Alignment Corpus—Notebook for PAN at CLEF 2015. In: [8]
22. Hopfgartner, F., Hanbury, A., Müller, H., Kando, N., Mercer, S., Kalpathy-Cramer, J., Potthast, M., Gollub, T., Krithara, A., Lin, J., Balog, K., Eggel, I.: Report on the Evaluation-as-a-Service (EaaS) Expert Workshop. SIGIR Forum 49(1), 57–65 (Jun 2015), http://sigir.org/forum/issues/june-2015/
23. Jayapal, A.: Similarity Overlap Metric and Greedy String Tiling at PAN 2012: Plagiarism Detection—Notebook for PAN at CLEF 2012. In: [10], http://www.clef-initiative.eu/publication/working-notes
24. Jayapal, A., Goswami, B.: Submission to the 5th International Competition on Plagiarism Detection. http://www.uni-weimar.de/medien/webis/events/pan-13 (2013), http://www.clef-initiative.eu/publication/working-notes, From Nuance Communications, USA
25. Khoshnavataher, K., Zarrabi, V., Mohtaj, S., Asghari, H.: Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation—Notebook for PAN at CLEF 2015. In: [8]
26. Kong, L., Han, Y., Han, Z., Yu, H., Wang, Q., Zhang, T., Qi, H.: Source Retrieval Based on Learning to Rank and Text Alignment Based on Plagiarism Type Recognition for Plagiarism Detection—Notebook for PAN at CLEF 2014. In: [7]
27. Kong, L., Lu, Z., Han, Y., Qi, H., Han, Z., Wang, Q., Hao, Z., Zhang, J.: Source Retrieval and Text Alignment Corpus Construction for Plagiarism Detection—Notebook for PAN at CLEF 2015. In: [8]
28. Kong, L., Qi, H., Du, C., Wang, M., Han, Z.: Approaches for Source Retrieval and Text Alignment of Plagiarism Detection—Notebook for PAN at CLEF 2013. In: [11]
29. Kong, L., Qi, H., Wang, S., Du, C., Wang, S., Han, Y.: Approaches for Candidate Document Retrieval and Detailed Comparison of Plagiarism Detection—Notebook for PAN at CLEF 2012. In: [10], http://www.clef-initiative.eu/publication/working-notes
30. Küppers, R., Conrad, S.: A Set-Based Approach to Plagiarism Detection—Notebook for PAN at CLEF 2012. In: [10], http://www.clef-initiative.eu/publication/working-notes
31. Meyer zu Eißen, S., Stein, B.: Intrinsic Plagiarism Detection. In: Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) Advances in Information Retrieval. 28th European Conference on IR Research (ECIR 06). Lecture Notes in Computer Science, vol. 3936 LNCS, pp. 565–569. Springer, Berlin Heidelberg New York (2006)
32. Mohtaj, S., Asghari, H., Zarrabi, V.: Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus—Notebook for PAN at CLEF 2015. In: [8]
33. Nourian, A.: Submission to the 5th International Competition on Plagiarism Detection. http://www.uni-weimar.de/medien/webis/events/pan-13 (2013), http://www.clef-initiative.eu/publication/working-notes, From the Iran University of Science and Technology
34. Oberreuter, G., Carrillo-Cisneros, D., Scherson, I., Velásquez, J.: Submission to the 4th International Competition on Plagiarism Detection. http://www.uni-weimar.de/medien/webis/events/pan-12 (2012),

http://www.clef-initiative.eu/publication/working-notes, From the University of Chile, Chile, and the University of California, USA
35. Oberreuter, G., Eiselt, A.: Submission to the 6th International Competition on Plagiarism Detection. http://www.uni-weimar.de/medien/webis/events/pan-14 (2014), http://www.clef-initiative.eu/publication/working-notes, From Innovand.io, Chile
36. Palkovskii, Y., Belov, A.: Submission to the 7th International Competition on Plagiarism Detection. http://www.uni-weimar.de/medien/webis/events/pan-15 (2015), http://www.clef-initiative.eu/publication/working-notes, From the Zhytomyr State University and SkyLine LLC
37. Palkovskii, Y.: "Counter Plagiarism Detection Software" and "Counter Counter Plagiarism Detection" Methods. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 67–68. Universidad Politécnica de Valencia and CEUR-WS.org (Sep 2009), http://ceur-ws.org/Vol-502
38. Palkovskii, Y., Belov, A.: Applying Specific Clusterization and Fingerprint Density Distribution with Genetic Algorithm Overall Tuning in External Plagiarism Detection—Notebook for PAN at CLEF 2012. In: [10], http://www.clef-initiative.eu/publication/working-notes
39. Palkovskii, Y., Belov, A.: Using Hybrid Similarity Methods for Plagiarism Detection—Notebook for PAN at CLEF 2013. In: [11]
40. Palkovskii, Y., Belov, A.: Developing High-Resolution Universal Multi-Type N-Gram Plagiarism Detector—Notebook for PAN at CLEF 2014. In: [7]
41. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: Braschler, M., Harman, D., Pianta, E. (eds.) Working Notes Papers of the CLEF 2010 Evaluation Labs (Sep 2010), http://www.clef-initiative.eu/publication/working-notes
42. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-Language Plagiarism Detection. Language Resources and Evaluation (LREV) 45(1), 45–62 (Mar 2011)
43. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection. In: Petras, V., Forner, P., Clough, P. (eds.) Working Notes Papers of the CLEF 2011 Evaluation Labs (Sep 2011), http://www.clef-initiative.eu/publication/working-notes
44. Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th International Competition on Plagiarism Detection. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) Working Notes Papers of the CLEF 2012 Evaluation Labs (Sep 2012), http://www.clef-initiative.eu/publication/working-notes
45. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 5th International Competition on Plagiarism Detection. In: Forner, P., Navigli, R., Tufis, D. (eds.) Working Notes Papers of the CLEF 2013 Evaluation Labs (Sep 2013), http://www.clef-initiative.eu/publication/working-notes
46. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
47. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B.: Overview of the 6th International Competition on Plagiarism Detection. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) Working Notes Papers of the CLEF 2014

Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2014), http://www.clef-initiative.eu/publication/working-notes
48. Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., Welsch, C.: ChatNoir: A Search Engine for the ClueWeb09 Corpus. In: Hersh, B., Callan, J., Maarek, Y., Sanderson, M. (eds.) 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12). p. 1004. ACM (Aug 2012)
49. Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: Fung, P., Poesio, M. (eds.) Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 13). pp. 1212–1221. Association for Computational Linguistics (Aug 2013), http://www.aclweb.org/anthology/P13-1119
50. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Huang, C.R., Jurafsky, D. (eds.) 23rd International Conference on Computational Linguistics (COLING 10). pp. 997–1005. Association for Computational Linguistics, Stroudsburg, Pennsylvania (Aug 2010)
51. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 1–9. CEUR-WS.org (Sep 2009), http://ceur-ws.org/Vol-502
52. Rodríguez Torrejón, D., Martín Ramos, J.: Detailed Comparison Module In CoReMo 1.9 Plagiarism Detector—Notebook for PAN at CLEF 2012. In: [10], http://www.clef-initiative.eu/publication/working-notes
53. Rodríguez Torrejón, D., Martín Ramos, J.: Text Alignment Module in CoReMo 2.1 Plagiarism Detector—Notebook for PAN at CLEF 2013. In: [11]
54. Rodríguez Torrejón, D., Martín Ramos, J.: CoReMo 2.3 Plagiarism Detector Text Alignment Module—Notebook for PAN at CLEF 2014. In: [7]
55. Rosvall, M., Bergstrom, C.: Maps of Random Walks on Complex Networks Reveal Community Structure. Proceedings of the National Academy of Sciences 105(4), 1118–1123 (2008)
56. Sanchez-Perez, M., Gelbukh, A., Sidorov, G.: Dynamically Adjustable Approach through Obfuscation Type Recognition—Notebook for PAN at CLEF 2015. In: [8]
57. Sanchez-Perez, M., Sidorov, G., Gelbukh, A.: A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014—Notebook for PAN at CLEF 2014. In: [7]
58. Sánchez-Vega, F., y Gómez, M.M., Villaseñor-Pineda, L.: Optimized Fuzzy Text Alignment for Plagiarism Detection—Notebook for PAN at CLEF 2012. In: [10], http://www.clef-initiative.eu/publication/working-notes
59. Saremi, M., Yaghmaee, F.: Submission to the 5th International Competition on Plagiarism Detection. http://www.uni-weimar.de/medien/webis/events/pan-13 (2013), http://www.clef-initiative.eu/publication/working-notes, From Semnan University, Iran
60. Shrestha, P., Maharjan, S., Solorio, T.: Machine Translation Evaluation Metric for Text Alignment—Notebook for PAN at CLEF 2014. In: [7]
61. Shrestha, P., Solorio, T.: Using a Variety of n-Grams for the Detection of Different Kinds of Plagiarism—Notebook for PAN at CLEF 2013. In: [11]
62. Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., Stein, B.: Overview of the PAN/CLEF 2015 Evaluation Lab. In: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 6th International Conference of the CLEF Initiative (CLEF 15). Springer, Berlin Heidelberg New York (Sep 2015)
63. Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic Plagiarism Analysis. Language Resources and Evaluation (LRE) 45(1), 63–82 (Mar 2011)

64. Stein, B., Meyer zu Eißen, S.: Near Similarity Search and Plagiarism Analysis. In: Spiliopoulou, M., Kruse, R., Borgelt, C., Nürnberger, A., Gaul, W. (eds.) From Data and Information Analysis to Knowledge Engineering. Selected papers from the 29th Annual Conference of the German Classification Society (GFKL 05). pp. 430–437. Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin Heidelberg New York (2006)
65. Suchomel, Šimon., Kasprzak, J., Brandejs, M.: Three Way Search Engine Queries with Multi-feature Document Comparison for Plagiarism Detection—Notebook for PAN at CLEF 2012. In: [10], http://www.clef-initiative.eu/publication/working-notes
66. Suchomel, Šimon., Kasprzak, J., Brandejs, M.: Diverse Queries and Feature Type Selection for Plagiarism Discovery—Notebook for PAN at CLEF 2013. In: [11]
67. Zarrabi, V., Rafiei, J., Khoshnava, K., Asghari, H., Mohtaj, S.: Evaluation of Text Reuse Corpora for Text Alignment Task of Plagiarism Detection—Notebook for PAN at CLEF 2015. In: [8]