Automatic Generation of Summary Obfuscation Corpus for Plagiarism Detection

Sabino Miranda-Jiménez[1], Efstathios Stamatatos[2]

[1]CONACYT / INFOTEC– Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Circuito Tecnopolo Sur 112, 20313, Aguascalientes, México, sabino.miranda@infotec.mx

[2]Department of Information and Communication Systems Engineering, University of the Aegean, Karlovasi, 83200, Samos, Greece, stamatatos@aegean.gr

Abstract: In this paper, we describe an approach to create a summary obfuscation corpus for the task of plagiarism detection. Our method is based on information from the Document Understanding Conferences related to years 2001 and 2006, for the English language. Overall, an unattributed summary used within someone else's document is considered a kind of plagiarism because the main author's ideas are still in a succinct form. In order to create the corpus, we use a Named Entity Recognizer (NER) to identify the entities within an original document, its associated summaries, and target documents. After, these entities, together with similar paragraphs in target documents, are used to make fake suspicious documents and plagiarized documents. The corpus was tested in plagiarism competition.

Keywords: corpus generation; plagiarism detection; obfuscation strategies

1 Introduction

The gigantic number of digital documents produced every day and the information available online, have made it easy to reuse data (sentences, excerpts, etc.) from others' work into one's own documents without citing the corresponding source of information; thus, plagiarism comes into the picture. Plagiarism is the reuse of someone else's ideas, processes, results or words without explicitly acknowledging the author's work and source [1].

In recent years, plagiarism detection has received much attention from the community in terms of published papers and systems developed, for example, PAN contests, in plagiarism detection task [2, 3]. In order to evaluate the systems developed, it is required corpus designed for this purpose. Traditionally, an intrinsic evaluation is conducted to evaluate the performance of systems [4, 5],

i.e., given a set of suspicious documents, a system must determine whether a whole document or sections of the document are plagiarized from other sources.

Thus, different corpora have been developed using, for example, obfuscation strategies such as author obfuscation [6], which consists in distorting the most frequent words for an author, replacing each word with one of their synonyms. Also, the use of paraphrasing was proposed in [7, 8], for example, using Wikipedia articles to creating the corpus of original and suspicious documents (fragments); in order to obfuscate fragments, it uses random strategies to shuffle words in the extracted fragments, in addition, Part-of-Speech features are used to preserve syntactic structure in fragments. A more sophisticated strategy uses SemEval dataset of semantic textual similarity [9], here, a pair of semantically similar sentences are used to create simulated plagiarism cases; both source and plagiarized fragments are constructed by SemEval dataset sentences.

In this work, we argue, in detail, the development of a corpus for plagiarism intrinsic evaluation used in evaluating systems of plagiarism detection [2]. The creation of the corpus is based on information (news dataset) from text summarization field. This dataset is usually used to evaluate the performance of summarization systems [10], i.e., the summary generated by systems is compared against abstractive/extractive summaries created manually by human experts. We chose abstractive summaries (no simple concatenation of sentences) related to this dataset because it could be considered as plagiarism of author's ideas. Our strategies are based on entities mentioned within news documents and theirs associated summaries and its similarity among target (suspicious) paragraphs/documents in order to mask the information.

In the following sections, we describe our approach in detail, in Section 2, the selection of documents from datasets used, and how the obfuscation strategies are applied. The resulting corpus and the performance of systems on our approach are presented in Section 3. Conclusions and future work are discussed in Section 4.

2 Obfuscation Approach

We propose a kind of plagiarism based on information comes from news and theirs associated summaries. In general, we consider that an unattributed summary used within a document of someone else is a kind of plagiarism, because the ideas of the author are still in a condensed form. For this work, we use summaries made by human experts who paraphrased the original news given in the dataset.

Roughly speaking, our method uses two news datasets: one for extracting summaries, and the other one for making the related suspicious documents. In the following subsections, we describe the creation of the corpus.

2.1 DUC Datasets

As we mentioned, the creation of the corpus is based on two news datasets from Document Understanding Conferences (DUC) in 2001 [10] and 2006 [11]. DUC competition provides these datasets for evaluating the performance of automatic text summarization systems; the datasets is for the English language. Generally, contests for evaluating the performance of systems reused their dataset for several years because of the cost of manual generation of datasets. DUC competition used essentially the same source of documents for DUC-2001 and DUC-2002, and another source of documents from DUC-2003 to DUC-2006.

In order not to choose the same documents (news) that we want to plagiarize and fake, we selected the datasets of DUC-2001 and DUC-2006 as our source of documents. The datasets are described as follows.

DUC-2001 news dataset comprises documents related to Wall Street Journal of years 1987-1992, AP newswire (1989-1990), San Jose Mercury News (1991), Financial Times (1991-1994), LA Times, and Foreign Broadcast Information Service (FBIS). For this dataset, four generic summaries for each document (news) with length of approximately 50, 100, 200, and 400 words were created manually by human assessors. The summary is considered as an abstractive document since no simple concatenation was implemented, i.e., the main ideas of the author, in a condensed form, are within the summary.

DUC-2006 news dataset comprises documents related to Associated Press newswire (1998-2000), New York Times newswire (1998-2000), and Xinhua News Agency (English version, 1996-2000).

Both datasets have similar topics but do not deal with the same news. Table 1 shows the number of documents selected as starting dataset. The news considered as original documents, was selected with at least a 400 word length. The associated summaries were summaries of a 100 word length, because this document length has enough information to deal with faking suspicious paragraphs. The target news was considered for documents with at least a 600 word length.

Table 1
Starting dataset for summary plagiarism task

Source	No. documents
DUC-2001: Original news	237
DUC-2001: Summaries per each news	2
DUC-2006: Target news	527

2.2 Similarity Measure

One of our strategies is using a similarity measure to identify similar paragraphs and similar documents. Therefore, we use an easy measure to calculate the similarity of two objects. In particular, we used the well-known measure, the Dice similarity or coefficient, that is simple but it has a good quality [12]. The Dice similarity is defined as follows, in equation 1.

$$Sim_{Dice}(X,Y) = 2\frac{X \cap Y}{|X| + |Y|} \quad (1)$$

In equation 1, let be X and Y documents, the similarity is between 0 and 1; 0 means no similarity, and 1 stands for maximum similarity. |X| means the cardinality of the document, and X represents the set of words of the document. For example, given two documents, X and Y, defined below, the Dice similarity for these documents is 0.1360; it means that 13.6% of X document is similar to Y document. In order to calculate the Dice similarity, we only used content words as elements to be compared, i.e., we discarded punctuation marks and words such as *a*, *to*, *on*, *or*, etc., known as Stop words. Stop words are considered that do not contribute to defining the content of the document [13].

X: *The British cattle industry was under siege, while many nations of the European Union were imposing or discussing bans on imports, fearing "mad cow" could be transmitted to humans*.

Y: *The controversial practice of feeding ground animal remains to pigs and poultry is to be outlawed across the European Union from January as part of a continent-wide effort to stamp out a rising wave of consumer panic over Mad Cow disease*.

2.3 Named Entity Recognition

Another key strategy is the entities mentioned in documents; entities identified are used by the main method. In news genre, entities are common, because data are facts about events, places, persons, dates, organizations, etc. In this genre, it could be easy to identify original documents and theirs associated summaries considering the occurrences of entities. Thus, in order to obfuscate the information using entities, we use a tool to extract them from texts.

The Stanford Named Entity Recognizer (NER) [14] is used to identify seven categories (entities) in documents: time, location, organization, person, money, percent, and date. The entity information and similarity measure are used to create fake documents, as well as fake paragraphs related to entities identified in documents. Figure 1 shows how the entities are identified, defined between XML-based tags, for example, **<LOCATION>** and **</LOCATION>**; and

<ORGANIZATION> and **</ORGANIZATION>**. The information between these tags is the entities to be distorted in target documents.

Investigators from the **<LOCATION>**United States**</LOCATION>** and <LOCATION>Egypt</LOCATION> will review part of the flight control system in the tail of **<ORGANIZATION>**Boeing**</ORGANIZATION>**'s 767 airplane as part of the investigation into the crash of <ORGANIZATION> EgyptAir</ORGANIZATION> Flight 990, the chairman of the <ORGANIZATION>National Transportation Safety Board</ORGANIZATION> said <DATE>Friday</DATE>. The disclosure comes just a couple days after the chairman of <ORGANIZATION>EgyptAir</ORGANIZATION> told a news conference in <LOCATION>Cairo</LOCATION> that something happened to the tail of the <ORGANIZATION>Boeing</ORGANIZATION> 767 that caused it to go into a near supersonic dive before the plane broke up and crashed into the sea. Safety board chairman <PERSON>Jim Hall</PERSON> said investigators from his agency and the <ORGANIZATION>Egyptian Civil Aviation Authority</ORGANIZATION> will examine the 767's elevator system, as well as perform a metallurgic examination of the plane's engine pylon components.

Figure 1
Entities identified by NER

2.4 Obfuscation Method

The creation of the summary obfuscation corpus is based on documents of two datasets of DUC competition. DUC-2001 dataset serves as original documents, i.e., these documents are the information to be plagiarized, and DUC-2006 dataset serves as target documents, these documents serve for two goals: the first one is to create plagiarized documents and the second one is to create suspicious documents, i.e., fake plagiarized documents, using the named entities and close documents related to the original documents according to their similarity.

In order to achieve the goals, our method includes three main stages: preprocessing of DUC datasets, candidate document selection, and data obfuscation.

2.4.1 Preprocessing of DUC Datasets

The first stage is selecting the documents from DUC datasets; initially, this information is used to measure the performance of text summarization systems. Thus, there are source documents and four or five associated summaries manually created by human experts.

As we mentioned, on one hand, the original documents were selected from DUC-2001 dataset. Each document was selected based on its length, the document size is greater than 400 words, and two associated summaries of 100 words were selected (see section 2.1); this set of documents we will refer to as *original documents*. Figure 2 illustrates an example of an original document and Figure 3 shows its associated summaries. We have two summaries for each original document.

On the other hand, the documents from DUC-2006 play the role of suspicious documents. Similarly, the documents selected were based on the length; the document size is greater than 600 words in order to have enough text to add fake paragraphs or plagiarized paragraphs; this set of documents we will refer to as *candidate documents*. In both cases, all HTML tags were removed and only the text body is used.

<DOC> **Coast Guard** and **Navy** aircraft and vessels today searched for a crewman missing from an F-14 jet fighter that plunged into the **Atlantic Ocean** off **North Carolina** while practicing combat maneuvers, killing his crewmate, officials said. Six people were injured in another F-14 crash Monday after two **Navy** aviators bailed out of their jet over an airfield in the San Diego suburb of El Cajon, sending it smashing into a hangar. And a pilot in Utah escaped injury today in a third military training flight in two days. The crash off Hatteras, N.C., occurred Monday afternoon 22 miles east of Oregon Inlet, the Navy said. A fishing boat picked up a crewman, who was pronounced dead. The identity of the dead aviator and his missing crewmate were not released pending notification of relatives. Five people, including the two Navy fliers, remained hospitalized today following the crash Monday morning in El Cajon 15 miles east of San Diego. The $35 million jet crashed upside down into hangars at Gillespie Field and exploded. The blaze ignited by the crash destroyed a hangar and an attached extension, but spared a nearby restaurant. Authorities said the two crewman tried to guide the jet to the runway at Gillespie Field before bailing out. Capt. Gary Hughes, commanding officer of Naval Air Station Miramar, said he was grateful there weren't more injuries, "particularly when you're this close to El Cajon. It's a very populated area." The jet passed within a mile of an elementary school. "I thought they were just doing tricks. And then we saw the parachutes," said Washington Moscuso, a sixthgrader at Ballantyne Elementary School. In the Atlantic accident, Lt. Cmdr. Mike John, a spokesman for the Navy's Atlantic Fleet air force in Norfolk, Va., said the plane was engaged in mock dogfights with another F-14 and an A-4 jet in restricted military airspace off the North Carolina coast. "It was flying a routine training mission," John said. The cause of the crash was not determined, officials said. The aircraft sank soon after impact, John said. The twin-engine supersonic fighter was attached to Fighter Squadron 143 at Oceana Naval Air Station in Virginia Beach, Va. In northern Utah today, an F-16A jet fighter crashed west of Hill Air Force Base after the pilot bailed out, a base spokeswoman said. The aircraft, assigned to Hill's 388th Tactical Fighter Wing, was on a routine training mission. Spokeswoman Silvia Le Mons-Liddle said the plane went down about 25 miles west of the base about 9:05 a.m. MDT. She said the crash site was in or near the Promontory Mountains, which are on a peninsula jutting into the Great Salt Lake, but she declined to be more specific. </DOC>

Figure 2
Example of an original document

<SUMMARY1> Today a Navy F-14 jet fighter plunged into the Atlantic off Hatteras, NC. One crewman is dead, the other missing. The plane was engaged in mock dogfight training when it crashed. It was attached to Fighter Squadron 143 at Oceana Naval Air Station in Virginia Beach, VA. Also today, an F-16A assigned to Hill Air Force Base's 388th Tactical Fighter Wing crashed in northern Utah. The pilot bailed out. These crashes followed Monday's crash of a Navy plane into a hangar at Gillespie Field at El Cajon, CA, a densely populated area. Five people, including the two crewmen remain hospitalized. **</SUMMARY1>**
<SUMMARY2> An F-14 jet fighter, attached to Fighter Squadron 143 at Oceana Naval Air Station in Virginia Beach, plunged into the Atlantic today off Hatteras, NC, while practicing combat maneuvers with another F-14 and an A-4 in restricted military airspace. One crewman is dead and another missing. Six people were injured Monday afternoon when another F-14 crashed into hangars at Gillespie Field in the San Diego suburb of El Cajon. The two Navy aviators ejected. An F-16 jet fighter assigned to Tactical Fighter Wing 388 in northern Utah crashed at 9:05am MDT today while on a routine training flight. The pilot ejected safely. **</SUMMARY2>**

Figure 3
Example of two associated summaries

2.4.2 Candidate Document Selection

The second stage consists in selecting the best document group from the candidate documents for each original document. In order to achieve this goal, we follow the next steps. First, for an original document is calculated the similarity with all documents in the candidate dataset. In order to do this, we used Dice similarity (equation 1) to identify the probable candidates to be obfuscated. The minimum threshold for similarity is 10 percent of the original document in order to guarantee at least a degree of similarity. Second, all nominee documents are ranked according to Dice similarity in descending order. After that, the top 10 documents are selected, as documents to be obfuscated (target documents). In addition, a nominee document could not be used more than ten times in order to give other documents the chance to be chosen. Two of the target documents are used for plagiarism and the remaining ones are for creating suspicious documents.

2.4.3 Data Obfuscation

The third stage consists in obfuscating the information of the target documents selected in the previous stage (Sec. 2.4.2). To achieve this goal, there are three steps: entity extraction, similar paragraph identification, and fake and plagiarized text insertion.

First, Stanford NER is applied to extract entities for each original document and its associated summaries, as well as the entities for each related target document. Second, paragraphs of target documents are selected according to most similar content by Dice similarity, and similar dispersion of entity types between the original document and the target document. Third, in order to insert the plagiarized summary, a random selection of the place in the target document is performed among the selected paragraphs in the previous step. In addition, two paragraphs are selected to be noisy areas, i.e., replacing the entities from the candidate paragraph with entities of the most similar paragraph from the original document, according to the content and the entity dispersion. The function of noisy areas is to mislead potential methods that take entities to identify plagiarism.

The entities extracted are used as round-robin approach, that is, a circular list, in order to continue replacing entities in the target paragraph until entities are exhausted. Figure 7 shows an example of plagiarized document with noisy areas. In the case of the generation of suspicious documents the entity identification and the replacement of named entities are applied in the same way. Fake, suspicious documents have a similar structure to plagiarized documents, without the plagiarized section. We can see in Figure 6, the original text, and we can see, in Figure 7, how the entities were replaced in the noisy area (entities are in bold), and the text is still readable, but it is obfuscated.

3 Results

The statistics of the resulting corpus are shown in Table 2. The corpus has 2370 documents. The corpus consists of two plagiarized documents and eight fake suspicious documents per each original document (237 documents). There are 496 plagiarized documents. A plagiarized document consists of the text, two noisy areas, and a plagiarized text (a summary), see Figure 7. The XML-based tags are only for informative purposes. Also, there are 1896 fake suspicious documents. A fake document consists of the text and two noisy areas, similar to the structure of Figure 7, without plagiarized section.

Table 2
Statistics of Obfuscation Summary Corpus

Source	No. documents
Original documents	237
Fake suspicious documents	1896
Plagiarized documents	474

Figure 4 and Figure 5 show the structure of annotations for fake suspicious documents and plagiarized documents respectively. The annotated documents are to identify what documents are plagiarized and what documents are only suspicious. In the case of a plagiarized document, there are key features such as the feature called *name* with value *plagiarism* that indicates that the current document has plagiarism; *source_offset* indicates the place where the plagiarism starts; *source_length* is the total of plagiarized characters; and *source_reference* indicates the file name. In the case of fake suspicious documents, this information is absent, see Figure 4. Note that XML-based tags in original documents, fake suspicious documents and plagiarized documents are only for informative purposes; these tags are not present in the final documents.

```
<document reference="suspicious-document00790.txt">
<feature name="about" authors="DUC2006" title="news" language="en" />
<feature name="md5" value="250487dd32850d9b89e1b094392609dc" language="en" />
</document>
```

Figure 4
Annotations for fake suspicious documents

```
<document reference="suspicious-document0010.txt">
<feature name="about" authors="DUC2001, DUC2006" title="news" language="en" />
<feature name="md5" value="9f6e52a04880aac50e92a3d300356a27" language="en" />
<feature name="plagiarism" type="artificial" obfuscation="high" this_language="en"
this_offset="3210" this_length="632" source_reference="source-document0001.txt"
source_language="en" source_offset="1" source_length="7224" />
```

Figure 5
Annotations for plagiarized documents

3.1 Evaluation of Systems

The corpus with the approach described in this paper was used in PAN competition for text alignment for plagiarism detection [2]. The task on this corpus is to determine whether the document contains plagiarized sections given a set of suspicious documents.

Table 3 shows the performance of systems using the summary obfuscation corpus and other two strategies used in the competition: Random obfuscation and Cyclic translation obfuscation, for more details of the implementation of the strategies see [2]. The performance of the systems was measured by *PlagDet* score. Basically, PlagDet is a measure that considers F1 score (harmonic mean of precision and recall) and a kind of normalization considering detections of passage cases with plagiarism and passages confirmed with plagiarism, this measure was designed for this purpose in PAN competitions, for more details of this measure see [2].

According to the performance of the systems with these datasets, it is hard for the systems to identify correctly the plagiarized documents as we can see in the low values obtained with our corpus (Summary Obfuscation) for all systems.

Table 3
Evaluation of text alignment systems related to plagiarism detection

Team	Random	Cyclic translation	Summary Obfuscation
Suchomel [16]	0.75276	0.67544	**0.61011**
Kong [17]	0.83242	0.85212	**0.43399**
R. Torrejón [18]	0.74711	0.85113	**0.34131**
Saremi [19]	0.65668	0.70903	**0.11116**
Shrestha [20]	0.66714	0.62719	**0.11860**
Gillam [21]	0.0419	0.01224	**0.00218**
Jayapal [22]	0.18148	0.18181	**0.05940**

In general, the performance for participating systems in plagiarism detection was weak. One system was a little higher than 60%, the remaining systems were below 45%. We notice intuitively, that the low performance of the systems is due to the strategies implemented and are not trivial, i.e., the plagiarized text is an abstractive summary made manually by human experts. An abstractive summary

is not a simple concatenation of sentences or excerpts from an original document; often, it is a complete paraphrasing of the text using several operations to abstract the text [15]. According to the results of the performance of systems, this approach presents great challenges to systems, when the text is a sort of plagiarized version of an author's ideas.

Figure 6

<TEXT>
Investigators from the **United States** and **Egypt** will review part of the flight control system in the tail of Boeing's 767 airplane as part of the investigation into the crash of **EgyptAir Flight** 990, the chairman of the **National Transportation Safety Board** said Friday. The disclosure comes just a couple days after the chairman of EgyptAir told a news conference in Cairo that "something happened" to the tail of the Boeing 767 that caused it to go into a near supersonic dive before the plane broke up and crashed into the sea. Safety board chairman Jim Hall said investigators from his agency and the Egyptian Civil Aviation Authority will examine the 767's elevator system, as well as perform a metallurgic examination of the plane's engine pylon components.
...
All 217 people aboard the **Boeing** 767-300 died when it plunged into the **Atlantic** off the **Massachusetts** coast on Oct. 31, about 30 minutes out of **New York's Kennedy Airport** on a night flight to Cairo. Investigators have found nothing in an analysis of the cockpit voice recorder that would point toward a bomb or a mechanical problem as the cause of its crash. Radio communication between the flight crew and air traffic controllers was routine, and at no time did a member of the crew advise controllers of either an emergency or a mechanical problem or concern. In addition, the plane's other black box, the flight data recorder, does not indicate there was an explosion or mechanical problem. That points to another cause, and the leading theory is that the plane was brought down by a deliberate act of the backup copilot. In his statement Friday, Hall blasted as "wrong" a published report this week that quoted unnamed government officials as saying a mechanical problem has all but been ruled out as the cause of the crash. "NTSB is disturbed to see that again this week unidentified sources were used as the basis of a news report purporting to have informed knowledge of our work," Hall said in a statement released late Friday afternoon. "As is often the case in these matters, the story was wrong. No hypothesis for the cause of this accident has been accepted, and the activities that I have outlined indicate that there is much that still needs to be done before a determination of cause can be reached." In November, with no evidence the crash was an accident, Hall was prepared to turn the investigation over to the FBI further fueling the theory of pilot suicide when the Egyptian government strenuously objected. Since then, the safety board has said little about how the investigation is going. Hall said substantial portions of the wings, tail, fuselage and an engine had been recovered. Hall said Friday that "no decision has been reached at this point whether further wreckage recovery will ultimately be necessary, and both agencies (the safety board and the Egyptian Civil Aviation Authority) agree that additional work needs to be be accomplished before a final decision can be made." Hall said both agencies also believe that "aircraft and operational system issues" must be investigated further.
</TEXT>

Example of a candidate document

Conclusions

We have described an approach for generating a summary obfuscation corpus to be used in plagiarism detection tasks. We used, as source of information, two datasets (DUC contest) from different years to avoid the same news. We considered the use of abstractive summaries within document as a case of plagiarism. We focused on, mainly, identifying entities and the dispersion of them through paragraphs, selecting similar documents and paragraphs to obfuscate the given summary into a plagiarized document; also twisting paragraphs replacing entities into noisy areas, for both plagiarized and fake documents. This approach was used in PAN competition for testing the performance of plagiarism detection systems.

<TEXT>
<NOISY_AREA>
Investigators from the **Atlantic** and **Hatteras** will review part of the flight control system in the tail of **Navy's** 767 airplane as part of the investigation into the crash of **NC Flight** 990, the chairman of the **Fighter Squadron** 143 said Monday. The disclosure comes just a couple days after the chairman of Oceana Naval Air Station told a news conference in Virginia Beach that " something happened" to the tail of the Hill Air Force Base 767 that caused it to go into a near supersonic dive before the plane broke up and crashed into the sea. Safety board chairman Jim Hall said investigators from his agency and the 388th Tactical Fighter Wing will examine the 767's elevator system, as well as perform a metallurgic examination of the plane's engine pylon components. The elevators are flat panels on the horizontal stabalizer of the tail that control up and down movements of the plane when the pilot pushes or pulls on the control stick. Hall's statement did not suggest that investigators suspect the elevator system played a role in the crash, and a Navy spokesman said such a review is typical in airline crash investigations. " The safety board is going through a deliberate and methodical process as they do on all their investigations, and Gillespie Field continues to support the investigation," said El Cajon safety spokesman John Dern . There have been no reports of problems with the 767's elevator system or the engine pylons, Dern said.
</NOISY_AREA>
<NOISY_AREA>
All 217 people aboard the **Coast Guard** 767-300 died when it plunged into the **Atlantic Ocean** off the **North Carolina** coast on Monday, about 30 minutes out of **San Diego's El Cajon** on a night flight to Utah. Investigators have found nothing in an analysis of the cockpit voice recorder that would point toward a bomb or a mechanical problem as the cause of its crash. Radio communication between the flight crew and air traffic controllers was routine, and at no time did a member of the crew advise controllers of either an emergency or a mechanical problem or concern. In addition, the plane's other black box, the flight data recorder, does not indicate there was an explosion or mechanical problem. That points to another cause, and the leading theory is that the plane was brought down by a deliberate act of the backup copilot. In his statement today, Gary Hughes blasted as " wrong" a published report this week that quoted unnamed government officials as saying a mechanical problem has all but been ruled out as the cause of the crash." Navy is disturbed to see that again this week unidentified sources were used as the basis of a news report purporting to have informed knowledge of our work," Washington Moscuso said in a statement released late Monday afternoon.
</NOISY_AREA>
" As is often the case in these matters, the story was wrong. No hypothesis for the cause of this accident has been accepted, and the activities that I have outlined indicate that there is much that still needs to be done before a determination of cause can be reached." In November , with no evidence the crash was an accident, Hall was prepared to turn the investigation over to the FBI further fueling the theory of pilot suicide when the Egyptian government strenuously objected. Since then, the safety board has said little about how the investigation is going. Hall said substantial portions of the wings, tail, fuselage and an engine had been recovered. Hall said Friday that " no decision has been reached at this point whether further wreckage recovery will ultimately be necessary, and both agencies (the safety board and the Egyptian Civil Aviation Authority) agree that additional work needs to be be accomplished before a final decision can be made." Hall said both agencies also believe that " aircraft and operational system issues" must be investigated further.
<PLAGIARIZED_TEXT1> Today a Navy F-14 jet fighter plunged into the Atlantic off Hatteras, NC. One crewman is dead, the other missing. The plane was engaged in mock dogfight training when it crashed. It was attached to Fighter Squadron 143 at Oceana Naval Air Station in Virginia Beach, VA. Also today, an F-16A assigned to Hill Air Force Base's 388th Tactical Fighter Wing crashed in northern Utah. The pilot bailed out. These crashes followed Monday's crash of a Navy plane into a hangar at Gillespie Field at El Cajon, CA, a densely populated area. Five people, including the two crewmen remain hospitalized.
</PLAGIARIZED_TEXT1>

Figure 7
Example of plagiarized document with two noisy areas

The results of the performance of systems showed that paraphrasing in a succinct way becomes great challenges to identify plagiarism.

As future work, we plan to apply our approach to multi-lingual and multi-document news, in the context of MultiLing competition [23, 24, 25]. In those datasets, there are same summaries in several languages such as Arabic, English, Greek, Hebrew and Spanish, but summaries are not a literal translation, they were

created by native speakers from original documents in the English language. In this sense, we could work on a corpus for cross-lingual plagiarism detection, based on abstractive summaries, i.e., a summary in a language "A" could be obfuscated in a language "B", a language "C", etc., considering the summary in "B" and "C" as a plagiarism of the summary in "A".

References

[1] A. Barrón-Cedeño, M. Vila, M. A. Martí, and P. Rosso: Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection, Computational Linguistics, Vol. 39, No. 4, 2013, pp. 917-947

[2] M. Potthast, M. Hagen, T. Gollub, M. Tippmann, J. Kiesel, P. Rosso, E. Stamatatos, and B. Stein: Overview of the 5th International Competition on Plagiarism Detection, in CLEF Conference on Multilingual and Multimodal Information Access Evaluation, 2013, pp. 301-331

[3] P. Rosso, F. Rangel, M. Potthast, E. Stamatatos, M. Tschuggnall, and B. Stein: Overview of PAN'16—New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation, in Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative, N. Fuhr, P. Quaresma, B. Larsen, T. Gonçalves, K. Balog, C. Macdonald, L. Cappellato, and N. Ferro, Eds. Berlin Heidelberg New York: Springer, 2016, pp. 332-350

[4] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso: An Evaluation Framework for Plagiarism Detection, in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, ser. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 997-1005

[5] M. Kuznetsov, A. Motrenko, R. Kuznetsova, and V. Strijov: Methods for Intrinsic Plagiarism Detection and Author Diarization, in CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal, K. Balog, L. Cappellato, N. Ferro, and C. Macdonald, Eds. Berlin Heidelberg New York: CEUR-WS.org, 2016, pp. 912-919

[6] M. Potthast, M. Hagen, and B. Stein: Author Obfuscation: Attacking the State of the Art in Authorship Verification, in Working Notes Papers of the CLEF 2016 Evaluation Labs, ser. CEUR Workshop Proceedings. CLEF and CEUR-WS.org, 2016, pp. 716-749

[7] M. Mansoorizadeh, T. Rahgooy, M. Aminiyan, and M. Eskandari: Author Obfuscation using WordNet and Language Models, in CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal, K. Balog, L. Cappellato, N. Ferro, and C. Macdonald, Eds. Berlin Heidelberg New York: CEUR-WS.org, Sep. 2016, pp. 939-946

[8] S. Mohtaj, H. Asghari, and V. Zarrabi: Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus, in CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France, L. Cappellato, N. Ferro, G. Jones, and E. San Juan, Eds. CEUR-WS.org, 2015

[9] E. Agirrea, C. Baneab, D. Cerd, M. Diabe, A. Gonzalez-Agirrea, R. Mihalceab, G. Rigaua, J. Wiebef, and B. C. Donostia: Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation, Proceedings of SemEval, 2016, pp. 497-511

[10] DUC-2001. (2001) The document understanding conference [Online] Available: http://duc.nist.gov/pubs.html#2001

[11] DUC-2006. (2006) The document understanding conference [Online] Available: http://duc.nist.gov/pubs.html#2006

[12] V. Thada and D. V. Jaglan: Comparison of Jaccard, Dice, Cosine Similarity Coefficient to Find Best Fitness Value for Web-retrieved Documents using Genetic Algorithm, International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2, No. 4, 2013, pp. 202-205

[13] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press, 2008

[14] J. R. Finkel, T. Grenager, and C. Manning: Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling, in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 363-370

[15] S. Miranda-Jiménez, A. Gelbukh, and G. Sidorov: Conceptual Graphs as Framework for Summarizing Short Texts, International Journal of Conceptual Structures and Smart Applications (IJCSSA), Vol. 2, No. 2, 2014, pp. 55-75

[16] Š. Suchomel, J. Kasprzak, and M. Brandejs: Diverse Queries and Feature Type Selection for Plagiarism Discovery, in CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, P. Forner, R. Navigli, and D. Tufis, Eds. CELCT, 2013

[17] L. Kong, H. Qi, C. Du, M. Wang, and Z. Han: Approaches for Source Retrieval and Text Alignment of Plagiarism Detectio, in CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, P. Forner, R. Navigli, and D. Tufis, Eds. CELCT, 2013

[18] D. Rodríguez Torrejón and J. Martín Ramos: Text Alignment Module in CoReMo 2.1 Plagiarism Detector, in CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, P. Forner, R. Navigli, and D. Tufis, Eds. CELCT, 2013

[19] M. Saremi and F. Yaghmaee: Submission to the 5th International Competition on Plagiarism Detection, PAN/CLEF 2013, CELCT, Semnan University, Iran, 2013, pp. 352-365

[20] P. Shrestha and T. Solorio: Using a Variety of n-Grams for the Detection of Different Kinds of Plagiarism, in CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, P. Forner, R. Navigli, and D. Tufis, Eds. CELCT, 2013

[21] L. Gillam: Guess Again and See if They Line Up: Surrey's Runs at Plagiarism Detection, in CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, P. Forner, R. Navigli, and D. Tufis, Eds. CELCT, 2013

[22] A. Jayapal and B. Goswami: Submission to the 5th International Competition on Plagiarism Detection, PAN/CLEF 2013, CELCT, From Nuance Communications, USA, 2013, pp. 352-365

[23] M. Elhadad, S. Miranda-jiménez, J. Steinberger, and G. Giannakopoulos: Multidocument Multilingual Summarization Corpus Preparation, Part 2: Czech, hebrew and spanish, in In MultiLing 2013 Workshop in ACL, Bulgaria, 2013, pp. 13-19

[24] G. Giannakopoulos: Multi-Document Multilingual Summarization and Evaluation Tracks in Acl 2013 Multiling Workshop, in Proceedings of the MultiLing 2013 Workshop on Multilingual Multidocument Summarization, Bulgaria, 2013, pp. 20-28

[25] G. Giannakopoulos, J. Kubina, F. Meade, J. Conroy, J. M. Bowie,J. Steinberger, B. Favre, M. Kabadjov, U. Kruschwitz, M. Poesio. Multiling 2015: Multilingual Summarization of Single and Multi-Documents, On-Line Fora, and Call-Center Conversations. Proceedings of SIGDIAL, Prague, 2015, pp. 270-274