Fusion of News Reports using Surface-Based Methods

Joel Azzopardi
Faculty of ICT
University of Malta,
Msida, Malta
joel.azzopardi@um.edu.mt

Christopher Staff
Faculty of ICT
University of Malta,
Msida, Malta
chris.staff@um.edu.mt

***Abstract*—Events occurring in the real world are covered by news reports from different sources. Each report generally contains information that is found in others, but may also contain unique information. To learn all the information about a particular event, a user will need to read all the different reports. This is a duplication of effort since most information will be repeated in the different reports.**

In our research, we attempt to fuse news reports about the same event into a single coherent document eliminating repetition but preserving all the information contained in the source reports using only surface-based methods. Information in each news report is represented by a set of entity relationship graphs. The graphs representing each report are then merged into a single graph whilst keeping track of the source sentences. The fused report is generated using the maximally expressive set of sentences – the sentences that carry most information about the entities and their relationships in the news report, and ensuring that all entities and relationships are expressed in the fused document.

Our Document fusion system was evaluated using a set of news reports downloaded from *MSNBC News* that cite their sources, and also using human evaluation. We show that our system is able to capture most of the information found across different source documents whilst maintaining readability.

***Keywords*-Document fusion; news; conceptual graphs; entity-relation graphs;**

I. Introduction

Within some time of an event happening in the real world, numerous news reports will appear on news sites on the World Wide Web describing that event. Moreover, as time passes, new reports will appear that give information on how that event developed and what its effects were. Since the news reports are written by different authors, each of which has his/her own sources and point of view, different reports on the same event may contain different pieces of information. News reports may agree with or contradict each other, or may contain contain unique pieces of information not found in the other reports.

A user may not have time to read all the available news reports, and therefore some information contained within these reports may remain hidden. A Document Fusion system assists the user by presenting to him/her a single document that contains all the information appearing within the source documents without repetition.

Document Fusion is related to Multi-document Summarisation [1], [2], but there are significant differences between these two areas. Given a set of related documents, the aim of multi-document summarisation is to produce a short summary of the important information contained within all the documents, whilst the aim of document fusion is to produce a document that contains all the information found within the individual documents, but without repetition.

Our aims in this research are to represent information in news reports conceptually to yield an underlying structure that can be used to compare news stories in order to identify what information is similar and novel; to be able to add new information into the conceptual structures as additional information about an event is published; and to present a single fused report to the user. All tasks are performed using only surface-based methods – we do not utilise knowledge bases and/or deep semantic processing. It is our opinion that the use of ontologies and/or semantic representations (as in [1]) will render the system domain dependent since then, it will require domain-specific pre-defined knowledge. We aimed for our system to be domain independent.

This paper proceeds as follows: in Section II we give an overview of related systems in literature; we describe our fusion system in Section III; Section IV describes how we performed our evaluation and presents the results obtained; and finally we present our conclusions in Section V.

II. Background

The aim of *Document Fusion* is to produce a document that contains all the required information without any repetition of facts [3], [1], [4]. Different news reports on the same event may agree with each other or may contradict each other. Therefore, for a user to get the most complete picture possible, he/she would need to read all the original reports since there is no guarantee that an existing single report presents the entire information about an event [4], [3], [1]. Fusion of different documents from across multiple sources also evens out the bias present in source reports [5].

Document Fusion is not a simple process due to the heterogeneity of sources. Each individual document has its own particular content, writing style and point of view – making it difficult to find correlations across documents, and

to identify repeated information [6], [3]. Such source differences also lead to the issue of the final fused document's readability [3].

The fusion process consists of [4], [3]:

- **Source Selection** – selection of the source documents to be fused together;
- **Source Representation** – representation of the different pieces of source information; and
- **Fusion** – merging the source information together to produce the fused document.

In a document fusion system, *Source Selection* can be performed by having an automatic clustering system producing clusters of related documents – e.g. clusters of news reports describing the same event.

The *Source Representation* phase is further sub-divided into: the segmentation of the source documents; and the construction of structured representations for each segment.

Source documents may be segmented into segments of different granularities, varying from coarse segments (such as paragraph segments) to very fine segments (e.g. phrases and individual words) [4]. [3] advises the use of paragraph segments – paragraphs are more context-independent and fusion of paragraph segments renders the resulting fused document more readable. On the other hand, [1] considers sentence segments to be too coarse since they may contain multiple themes. [1] performs fusion using phrase segments, but then requires sentence regeneration to build the final fused report. Language generation systems are typically implemented for limited domains, as they require rich semantic representations [1]. Midway between these two 'extremes', [6] advocates the use of sentence segments.

The most common approach for segment representation is by using the *Bag-Of-Words* representation, whereby each segment is represented by the words it contains [5], [7]. On the other hand, [4], [1] and [8] utilise graph-like structures to represent their text segments: [4] builds document trees representing the structures of the phrases and sentences within the documents; [1] uses the *DYSNT* structure whereby a sub-graph is constructed for each phrase – a phrase is considered to be a verb with two nouns, and a graph node is built for each of these components; and [8] utilises dependency tree structures to represent the dependencies between the words in sentences.

The *Fusion* phase entails the fusion of the source information's representation and the selection of those information segments that will ultimately form part of the final fused document. [1] mentions that relationships between different segments can be identified *semantically* – using knowledge bases, or *statistically* – using statistical methods. *Semantic* methods can only be used in constrained domains whilst *statistical* methods may be utilised in arbitrary domains.

[4] lists various ways in which inter-segment relationships can be established, namely: using *Lexical Distance*, *Lexical Chains*, *Information Extraction*, and *Linguistic Pattern Matching*. [3] utilises the *subsumption/entailment* defined in [4] to find if one segment contains more information than another, and then using these entailment scores in the selection and ordering of paragraph segments to be included in the fused document. [1] identifies segments' relationships by finding common nodes within the DSYNT structures representing each phrase. Similarly, [8] attempts to align the syntactic trees of sentences from different documents.

After identifying the sets of segments that contain similar information, document fusion systems need to select the representative segments that will form part of the final fused document. To remove source bias, [8] and [1] generate a new sentence for each set of similar sentences. Sentence ordering is then performed by time-stamping each segment set with the earliest publication time of its member sentences/phrases. In contrast to this, [3] selects a base document, and then attempts to replace each paragraph in the base document by searching for another paragraph (from the documents other than the base document) which maximally entails the paragraph in question. This procedure ensures that the order of the fused report always follows that found in the base document.

III. Methodology

In the previous section, we described how the fusion process consists of *Source Selection*, *Source Representation* and *Fusion*. The *Source selection* phase in our news report fusion system is undertaken by a specific categorisation component that we have developed that clusters news reports into event-centric clusters – i.e. the news reports in each cluster describe the same event.

The source representation phase is further subdivided into the segmentation of the source documents and the construction of a logical structure for each resulting segment. We segment source news reports into sentence segments – the same granularity as used in [9], [2], [6], [10]. In our opinion, paragraph segments would be too coarse, and would not allow the selection of the optimal set of segments in such a way that all information is represented without any repetition. On the other hand, using finer segments such as phrases (as in [1]), would necessitate the use of sentence-regeneration and semantic processing.

After the source reports have been segmented into sentence segments, each sentence segment is represented using one or more entity-relation structures. In our opinion, the ideal means of information representation are conceptual graphs (described in [11], [12]) as these are able to represent all the knowledge necessary for people to comprehend a language or a text – namely *Lexical Knowledge*, *Syntactic Knowledge*, *Semantic Knowledge* and *Episodic Knowledge*. However, surface-based methods on their own are not able to decipher and represent *Semantic Knowledge* and *Episodic Knowledge*. Since in our approach we utilise only surface-based techniques to retain domain independence, we use

entity-relation graphs – these may be considered as a simplified form of conceptual graphs.

Each entity-relationship structure consists of: two entities (noun entities) and a relationship name (verb entity) describing the relationship between these two entities; or of a noun entity and a verb entity describing an intransitive action by that noun entity. These structures are quite similar to the *DSYNT* structures described in [1], and to the graphical structures used in [13] and [14]. All these structures consist of concepts built from noun entities, and verb entities describing the relationships between these concepts. Our entity-relationship structures differ from the *DSYNT* structures in that the terms forming a complete noun or verb phrase are clustered together in a single node within our system, and are not considered as separate nodes. Moreover, our system does not utilise *WordNet* (as in [14]) to perform sense disambiguation and identify synonyms. Also, our system does not represent knowledge with different levels of abstraction as in [13]. In our system, noun and verb entities are identified using heuristic rules based on Part-Of-Speech (POS) tags – this approach was also applied in [13], [14] and [15]. We use Brill's Part of Speech tagger [16] to tag the source documents' text.

In order to retain domain independence, we do not utilise any *semantic* methods to identify relationships between different entities. The discovery of relationships performed by our system is similar to the discovery of *named relationships* employed by [15], and also to the relation discovery described in [14]. Relationships are found by matching POS patterns between different noun entities and a term (verb entity) describing the relationship. The discovered relationships in our system may be a *binary relation* where the verb entity involved is transitive and the relation is defined between two noun entities (e.g. in the phrase '*John kicked Mary*', we have the relation '*kicked*' between '*John*' and '*Mary*'); or a *unary relation* where there is only one noun entity object involved since the verb 'defining' the relation is intransitive (e.g. in the phrase '*John died*', we have the 'relation' '*died*' and only the noun entity '*John*' is involved).

After all the segments within a news report cluster have been represented as entity-relationship structures, these structures are compared to identify co-referent entities and similar relationships across different structures. The identification of co-referent entities from across different entity-relation structures is based on the amount of overlapping salient terms within the different entity objects. This approach is similar to an approach described in [17] where key phrases that contain similar 'centre' words are clustered together. Salient terms within each entity object are found by calculating *TF.IDF* weights of each term in the noun and verb entities and normalising these weights by dividing these weights with the maximum weight in each noun/verb entity. Those terms that have a *normalised* weight of 0.5 or greater are considered to be the salient tokens for that entity object.

After the clustering of co-referent entities and relationship names, those relationships that are conveying the same information are clustered together. Similar relationships are those that have 2 co-referent entities, or that have 1 co-referent entity and a similar relationship name. As a result of the entity-relationship structures' clustering, each news report cluster should ideally consist of a set of entity-relationship clusters each conveying 'unique' information. Therefore, the final fused report should contain a reference to each of these entity-relationship clusters.

We construct the final fused document by searching for the optimal set of sentences from the source documents whereby if possible, all entity-relationship clusters are represented, but only once. This task is not trivial since many sentences represent multiple entity-relationship structures. The sentences are output in the final fused report using the natural ordering approach [9] – each entity-relation is time-stamped with the earliest time it became known, and the sentences in the fused report are then ordered according to the earliest time stamp of the entity-relation they reference.

To enhance readability, each fused report is sub-divided into sections according to the date and time the different pieces of information became known, and also into paragraphs using *TextTiling* [18].

IV. Evaluation

Document Fusion systems can be evaluated in two different manners [3]: *intrinsically* where evaluation is done (usually by humans) on the quality of the fused document itself; and *extrinsically* where the usefulness of the fused document in the completion of a different task is evaluated. Intrinsic evaluation of document fusion systems has the added difficulty that there is no existing gold standard for such evaluations [1], [3]. For the evaluation of our system, we use a suggestion put forward by [3] to use news reports that cite their sources. We also utilise human input in the evaluation of our fused news reports.

A. Automated Evaluation

For this part of the evaluation, we used news reports downloaded from *MSNBC News* website[1] since a number of news reports on this portal cite their sources. We parsed out the sources from the downloaded news reports and discarded those reports that do not cite at least two sources. We then searched for the source news reports using *Google News*[2] – *Google News* allows the search of news report from user-specified sources. We searched for the source news reports by specifying the report title as the search text, and specifying also the cited source to limit search results from only that source. We then gradually reduced search terms until a result is found, or until the search terms have been

[1] *http://www.msnbc.msn.com*
[2] *http://news.google.com*

reduced by half in which case the search for the source report is abandoned.

Since the described method may return unrelated or minimally related source reports, each downloaded source report was compared with the corresponding *MSNBC News* report and in cases where the *Cosine similarity* between them (each report was represented using the *Bag-Of-Words* model, and all terms were stemmed, filtered from stop words and weighted using *TF.IDF*) was less than a pre-defined threshold, the downloaded source report was discarded.

We then proceeded with our evaluation using those *MSNBC* news reports that had their full complement of cited source reports located and downloaded – the final corpus used consisted of 79 *MSNBC News* reports that had their full complement of source reports (generally 2 or 3). We had our system perform fusion of each set of source reports, and each generated fused report was compared to the corresponding *MSNBC News* report (this was considered to be the 'ideal' fused report). We acknowledge the fact that this evaluation corpus has its errors – there is no proof that the downloaded source reports were the actual source reports used in the construction of the *MSNBC* report; the *MSNBC* report may not contain all the information found in the source reports; and the *MSNBC* report author may have used information from other uncited sources. However, we had no better evaluation corpus at our disposal.

For the first part of the automated evaluation, we evaluated the average sentence-to-sentence similarity between each *MSNBC* report and the corresponding produced fused report[3]. Then we compared this similarity to the average sentence-to-sentence similarity between each *MSNBC* report and the corresponding source reports. This allowed us to evaluate whether the produced fused report is more similar to the 'ideal' fused report, and contains more information than the individual source reports. We did not measure information overlap by direct string matching since most probably the ideal fused report has been constructed by humans and did not involve direct sentence extraction. Also two sentences may be conveying the same information but utilising different terms. Table I shows the results obtained.

The results in Table I show that on average, the ideal fused report has a higher information overlap with the produced fused report than with the source reports. This means that by the fusion process, the source reports are rendered closer to the ideal fused report, and therefore fusion is not being done in vain. In this evaluation, we had some cases where the most similar source report has a higher information overlap with the ideal fused report than the produced fused report. This is most probably caused by certain short-comings in the corpus used where the source reports used may not have been used at all in the construction of the corresponding *MSNBC News* report.

[3]The sentence-to-sentence similarity is calculated using *cosine similarity* after representing each sentence with the bag-of-words it contains, performing suffix stripping, stop-word removal and weighting each term using *TF.IDF*. In our evaluation we considered two sentences having a similarity of at least 0.1 between them as as conveying similar information.

Table I
AVERAGE SENTENCE-TO-SENTENCE SIMILARITY BETWEEN 'IDEAL' FUSED REPORTS, THE PRODUCED FUSED REPORTS, AND THE SOURCE REPORTS

Description	**Average Similarity**
Mean sentence-to-sentence similarity between 'ideal' and produced fused reports	0.3522
Minimum sentence-to-sentence similarity between 'ideal' and produced fused reports	0.01333
Maximum sentence-to-sentence similarity between 'ideal' and produced fused reports	0.7679
Mean sentence-to-sentence similarity between 'ideal' fused report and source reports	0.3503
Minimum sentence-to-sentence similarity between 'ideal' fused report and source reports	0.0120
Maximum sentence-to-sentence similarity between 'ideal' fused report and source reports	0.7803

In a second experiment, we evaluated the representation of each sentence in the source reports within the produced fused report, and within the 'ideal' fused report. We performed this test by finding for each sentence in each source report, the most similar sentence in the produced fused report. If the most similar fused report sentence had a similarity with the source report sentence that was less than the predefined threshold, then that source report sentence was considered to be unrepresented within the fused report. This allows us to evaluate the coverage of information in the source reports within the produced fused report. We did the comparison on a sentence by sentence basis, since the fusion process is performed on sentence segments. The results quantifying the representation of the sentences in the source reports within the corresponding fused reports are presented in Table II.

Table II shows that an average of 96% of the sentences in the source reports are represented in the produced fused reports. On the other hand, only 68% of the source reports' sentences are represented in the 'ideal' fused reports. As we mentioned previously, one must bear in mind the possibility that the source reports used may not have been the actual sources used during the construction of the 'ideal' fused reports. Also, it may imply that the MSNBC News reports summarise some of the information contained in the source reports rather than faithfully fusing all of the source reports.

We also compared the lengths of the produced and 'ideal' fused reports with the lengths of the corresponding source reports to ensure that the fused reports are not simply a concatenation of the different source reports. A fused report with length equal to the sum of the source reports' lengths defies the scope of document fusion. Table III shows a comparison of the length (in number of words) of the fused news reports with the sum of the corresponding source reports' length.

Table II
REPRESENTATION OF SOURCE SENTENCES IN THE FUSED REPORTS

Description	Inclusivity
Mean Source Report Sentence Representation in Produced Fused Report	96%
Mean Source Report Sentence Representation in 'Ideal' Fused Report	68%
Minimum Source Report Sentence Representation in Produced Fused Report	75%
Minimum Source Report Sentence Representation in 'Ideal' Fused Report	10%
Maximum Source Report Sentence Representation in Produced Fused Report	100%
Maximum Source Report Sentence Representation in 'Ideal' Fused Report	100%

Table III
COMPARISON OF REPORT LENGTHS (VALUES ARE GIVEN AS A PERCENTAGE OF THE TOTAL SOURCE REPORTS' LENGTH)

Description	Length
Mean Produced Fused Report Length	85%
Minimum Produced Fused Report Length	48%
Maximum Produced Fused Report Length	100%
Mean 'Ideal' Fused Report Length	62%
Minimum 'Ideal' Fused Report Length	6%
Maximum 'Ideal' Fused Report Length	193%

One can note that on average each produced fused report has a length of 85% of the source reports' total length, whereas each 'ideal' fused report has an average length of 62% of the sources' total length. However, when one compares these values to the ones quantifying the representation of source report information in the fused reports (Table II), one can note that the produced fused reports represent more of the information found in the source reports.

In three cases (approx. 4%), the produced fused report is equal in length to the sum of different source reports' lengths. This means that the fused report contains all the sentences found in the different source lengths, and that no overlap was found. Although this indicates that our fusion system might not be identifying overlap of information well enough, its underlying cause might also be that the source reports are totally different from each other.

B. Human Evaluation

The evaluation described in the previous section evaluated the inclusion of information content in the fused reports but they do not evaluate the readability and coherence of the fused reports. Therefore, we had a number of anonymous users provide readability ratings for fused news reports on current events from a web portal set up for this purpose. Users were presented with the fused news reports, links to the source news reports, and an online evaluation form for each fused news report. Users were asked to fill in the evaluation form with their ratings for: *fact completeness* – whether the fused news reports are covering all the facts in the source reports; *fact ordering* – whether the ordering of facts in the produced fused news reports are correct; and *redundancy* – whether the fused news reports contained repeated information. For each of the above rating, users were asked to provide a score between 1 and 5. A document scoring 5 in each rating means that it is covering all the facts in the source report, its facts are ordered correctly, and it has no redundancy. 25 anonymous evaluators responded to our invitation for the evaluation of our system and submitted valid evaluation records. The average scores obtained for each rating are shown in table IV.

Table IV
HUMAN RATINGS OBTAINED FOR THE PRODUCED FUSED NEWS REPORTS

Rating	Average Score
Fact Completeness	3.00
Fact Ordering	3.36
Redundancy	2.67

The results obtained indicate the our system performs fairly well – the results are neither very good, and nor poor. However, one must keep in mind that these anonymous ratings were provided by the general public – the evaluators do not consider that the fused documents were produced automatically without human intervention, and that it is very hard for an automatic system to generate fused documents that are equivalent in quality to ones written by humans. The results also indicate that our system needs to be improved to better identify and reduce redundancy of information in the fused document.

V. CONCLUSION

In this paper, we have shown how one can utilise surface-based approaches to represent news reports using entity-relation structures in order to allow the system to identify what information is similar and novel, and to fuse related reports into a single coherent report with minimal or no repetition but preserving all the information contained in the source reports. These approaches do not utilise knowledge bases and/or deep semantic analyses, and it is our opinion that they can be applied with relative ease in other domains apart from news. Since the described approaches utilise only shallow techniques, document fusion is performed quite rapidly without the need of significant resources. This means that the document fusion system described can be implemented as part of an operational system that fuses breaking news reports to present to interested readers complete descriptions of events using all the available information.

Our evaluation results show that on average there is a higher overlap of information between the 'ideal' fused report and produced fused report than between the 'ideal' fused report and the corresponding source reports. This means that the fusion process we have developed is useful. Results also show that on average 96% of all the sentences in the source reports are represented in the corresponding

produced fused report – i.e. the fusion process captures the majority of information in the source reports. Moreover, the fact that the average length of each fused report is 85% of the total length of the corresponding source reports indicates that the fusion process is not just a concatenation of the different documents.

Results obtained from the human evaluation indicate that further improvements are required to have better identification of repeated information across different source reports. Such redundancies stem mainly from the use of different terms and varied sentence structures across the different source reports. The identification of synonyms and the discovery of broader relationships between different entities will probably help in reducing repeated information. A possible improvement to our system can be the identification of *unnamed relationships* as performed in [15] and [7]. *Unnamed relationships* are statistically significant co-occurrence relationships that indicate that two entities are more or less strongly related but the nature of their relationship is unknown.

Further research can also be performed in the identification of contradictions across the different source reports. In our fusion approach, contradictory pieces of information are considered as different pieces of information, resulting in a contradiction within the generated fused report. Ideally, if two source reports are giving contradictory information, the contradiction should be highlighted and the sources of the different versions of information should be cited.

REFERENCES

[1] R. Barzilay, K. R. McKeown, and M. Elhadad, "Information fusion in the context of multi-document summarization," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1999, pp. 550–557.

[2] S. Harabagiu and F. Lacatusu, "Topic themes for multi-document summarization," in *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 2005, pp. 202–209.

[3] C. Monz, "Document fusion for comprehensive event description," in *Proceedings of the workshop on Human Language Technology and Knowledge Management*. Morristown, NJ, USA: Association for Computational Linguistics, 2001, pp. 1–8.

[4] D. R. Radev, "A common theory of information fusion from multiple text sources step one: cross-document structure," in *Proceedings of the 1st SIGdial workshop on Discourse and dialogue*. Morristown, NJ, USA: Association for Computational Linguistics, 2000, pp. 74–83.

[5] Y. Zhai and M. Shah, "Tracking news stories across different sources," in *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*. New York, NY, USA: ACM Press, 2005, pp. 2–10.

[6] Y. Zhang, C.-H. Chu, X. Ji, and H. Zha, "Correlating summarization of multi-source news with k-way graph bi-clustering," *SIGKDD Explor. Newsl.*, vol. 6, no. 2, pp. 34–42, 2004.

[7] D. T. Bollegala, Y. Matsuo, and M. Ishizuka, "Relational duality: unsupervised extraction of semantic relations between entities on the web," in *WWW '10: Proceedings of the 19th international conference on World wide web*. New York, NY, USA: ACM, 2010, pp. 151–160.

[8] R. Barzilay and K. R. McKeown, "Sentence fusion for multidocument news summarization," *Comput. Linguist.*, vol. 31, no. 3, pp. 297–328, 2005.

[9] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies," in *NAACL-ANLP 2000 Workshop on Automatic summarization*. Morristown, NJ, USA: Association for Computational Linguistics, 2000, pp. 21–30.

[10] Q. L. Israel, H. Han, and I.-Y. Song, "Focused multi-document summarization: human summarization activity vs. automated systems techniques," *J. Comput. Small Coll.*, vol. 25, no. 5, pp. 10–20, 2010.

[11] J. F. Sowa, "Semantics of conceptual graphs," in *Proceedings of the 17th annual meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1979, pp. 39–44.

[12] J. F. Sowa and E. C. Way, "Implementing a semantic interpreter using conceptual graphs," *IBM J. Res. Dev.*, vol. 30, no. 1, pp. 57–69, 1986.

[13] R. Richardson and E. A. Fox, "Using concept maps in digital libraries as a cross-language resource discovery tool," in *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA: ACM Press, 2005, pp. 256–257.

[14] K. Rajaraman and A.-H. Tan, "Knowledge discovery from texts: a concept frame graph approach," in *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*. New York, NY, USA: ACM Press, 2002, pp. 669–671.

[15] R. Byrd and Y. Ravin, "Identifying and extracting relations in text," in *NLDB 99 – 4th International Conference on Applications of Natural Language to Information Systems*, Klagenfurt, Austria, 1999.

[16] E. Brill, "A simple rule-based part of speech tagger," in *Proceedings of the third conference on Applied natural language processing*. Morristown, NJ, USA: Association for Computational Linguistics, 1992, pp. 152–155.

[17] S. Sakurai and A. Suyama, "Rule discovery from textual data based on key phrase patterns," in *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*. New York, NY, USA: ACM Press, 2004, pp. 606–612.

[18] M. A. Hearst, "Texttiling: segmenting text into multi-paragraph subtopic passages," *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, 1997.