News Search Using Discourse Analytics

Paul Thompson, Raheel Nawaz, Ioannis Korkontzelos, William Black, John McNaught and Sophia Ananiadou
National Centre for Text Mining, School of Computer Science
University of Manchester
Manchester, UK
{paul.thompson, ioannis.korkontzelos, william.black, john.mcnaught, sophia.ananiadou}@manchester.ac.uk,
raheelnawaz78@hotmail.com

***Abstract*— The vast numbers of digitised documents containing historical data constitute a rich research data repository. However, computational methods and tools available to explore this data are still limited in functionality. Research on historical archives is still largely carried out manually. Text mining technologies offer novel methods to analyse digital content to identify various types of semantic information in these documents and to extract them as semantic metadata. Methods range from the automatic identification of named entities (e.g., people, places, organisations, etc.) to more sophisticated methods to extract information about *events* (e.g., births, deaths, arrests, etc.), allowing users to greatly increase the specificity of their search. We have created an extended model of event interpretation to allow searches to be refined based on various discourse facets, including isolating definite information about events from more speculative details, distinguishing positive and negative opinions and categorising events according to information source. We present ISHER as an example of a multi-faceted, semantically oriented system for searching news articles from the New York Times, dating back to 1987. We explain how our extended event interpretation model can enhance search capabilities in systems such as ISHER, including the identification of contrasting and contradictory information in news articles.**

Keywords—semantic metadata; social history, discourse analysis; text mining; events; event interpretation; event-based search

I. INTRODUCTION

The digital information era has made vast and continually growing amounts of data available in digital form. Amongst these data are news archives, which provide an important source of information for researchers, social historians or members of the public interested in events captured in news stories. Web-based search systems that specifically target news articles (such as Google News[1] and Yahoo! News Search[2]) are readily available. However, despite the wealth of useful information locked away within these archives, their sheer size means that interested parties may struggle to unlock their full potential. Users of such search systems need effective means to isolate relevant information from the mountain of irrelevant information. Based on this, both Google and Yahoo offer advanced search facilities, providing various additional criteria for filtering search results. These include the date of publishing, publisher, location and category of the article.

A drawback of the above filtering criteria is that they are mostly based on high-level article features. The only means of placing restrictions on the *content* of the retrieved articles is usually to specify keywords. However, keywords do not allow aspects of the *meaning* of the articles to be taken into account. Consider that a user wants to find information about crimes that have occurred in the town of Sandwich. The keywords *crime* and *Sandwich* will return many irrelevant documents in which *Sandwich* refers to the food rather than the town. Links between terms can also be important. For example, if a researcher is interested in finding out information about lethal atrocities carried out by Saddam Hussein in Iraq, she may enter the search terms *Saddam Hussein, kill* and *Iraq*. Such a search is, however, likely to retrieve irrelevant stories as well as relevant ones, since, e.g., the search terms used could also be found in articles mentioning attempts on Saddam's own life. Such limitations in the expression of search criteria in popular search engines mean that, despite the high-level filtering mechanisms provided, users may still have to spend a long time sifting through search results in order to find articles of relevance to them.

In this paper, we examine how s*emantic-based searching* can enable users to more easily satisfy their information needs. By allowing search criteria to take into account various aspects of the meaning of the articles in the archive, retrieved results can be much more closely related to users' areas of interest. We describe how the New York Times (NYT) annotated corpus [1] constitutes a first step towards improving search, since it contains semantic metadata for around 1.8 million articles published in the NYT, over a period of 20 years. We subsequently look at how the application of text mining methods can help to improve search capabilities by making sense of the unstructured knowledge contained within texts. Such methods facilitate, e.g., the automatic identification and extraction of named entities (e.g., people, locations, organisations, etc.), together with relationships between them. The extracted details can be added as semantic metadata to the articles in the archive. By creating systems that are able to search this metadata, we can place more sophisticated restrictions on the content of the articles to be retrieved. Stimulated by numerous challenges to encourage researchers to develop increasingly sophisticated methods of extracting semantic information from various types of texts, including newspaper articles (e.g., [2, 3]), text mining

[1] https://news.google.com/
[2] http://news.search.yahoo.com/

This work has been funded by UK Joint Information Systems Committee (JISC) as part of the ISHER project, awardee of the Digging into Data Challenge.

technology has now reached a sufficient level of maturity to allow the extraction of structured information about *events* described in text (e.g., social unrest, strikes, arrests, convictions, etc.), including various details about these events (e.g., how individuals or organisations are involved in them).

We look at how the recognition of discourse-based features of events (e.g., negation, speculation, opinions towards them, information source, etc.) can allow users to further restrict their searches, and to discover and compare contrasting and conflicting information that has been reported for the same event. We examine how a richly annotated resource containing news articles (the ACE 2005 corpus [4]) can act as a basis for training systems to extract event and discourse-based metadata. We discuss some weaknesses of the original ACE resource, in terms of capturing comprehensive discourse information, according to which we propose an updated annotation model, and use it to create a more suitable training resource.

As a concrete demonstration of how rich semantic metadata can significantly enhance search capabilities, we present our multi-faceted, integrated semantic search system (ISHER) that uses text mining methods to extract semantic metadata from digitised historical newspaper archives of the NYT (1987 - 2007), based on training on the ACE 2005 corpus.

II. The Need for Semantic Search

To motivate the need for semantically-based search systems, let us return to the scenario of finding information about people killed by Saddam Hussein in Iraq. The researcher may start by entering the following search terms: *"Saddam Hussein", "kill"* and *"Iraq"*. Whilst such a query will retrieve many relevant articles, it will also match many irrelevant ones. Furthermore, it will also fail to retrieve further articles that *are* relevant. Some reasons are as follows:

- In the articles returned, search terms may not be related in the way intended by the user, or even occur in the same sentence. Therefore, references to *killing* mentioned in the retrieved article may not involve Saddam at all.
- The search term *Saddam Hussein* does not necessarily refer to a person. It may also refer to a town in Sri Lanka.
- There are multiple ways in which *Saddam Hussein* could be referred to in text, e.g., *Hussein*, *the dictator*, *the leader of Iraq*, etc. However, the interpretation of some of these is context-dependent. For example, *Hussein* is a common name and is shared by other well-known figures (e.g., King Hussein of Jordan). Additionally, several countries have dictators, and Iraq (like any other country) has had several leaders.
- A range of verbs or nouns may be used to denote relevant events, e.g., *murder, execution,* etc.
- Words and phrases appearing in the discourse context of the event may significantly alter the interpretation, e.g. *Hussein may have killed X* is different to *Hussein has killed X*.

The above issues imply that simple, non-semantic keyword search is not adequate for retrieving relevant information, without formulating several queries that try to account for possible variations in the expression of information in text.

A way of ensuring that search terms are related (in the desired manner) is to formulate longer query terms, such as "*Saddam Hussein kill*", which will retrieve only those articles in which the verb *kill* (or one of its inflections, like *killed* or *kills*) directly follows the name *Saddam Hussein,* thus identifying *Saddam* as the killer. This method would ensure that a higher percentage of the search hits would contain relevant events than if separate query terms were used. However, the many variations in the way that events can be expressed in natural language mean that such a fixed query string is likely to miss more relevant events than it actually retrieves. This is because, in specific articles, additional words and phrases may be inserted between "*Saddam Hussein*" and "*kill*", e.g. *yesterday* or *mercilessly*. Furthermore, this query restriction (i.e., where the word *kill* directly follows the word *Saddam Hussein*) does not cater for sentences written in the passive voice, e.g., *X was killed by Saddam Hussein*.

Taking the above into account, our semantically-based search engine offers the following features:

1. It allows specification of the semantic types of search terms, e.g., so that articles mentioning *Saddam Hussein* are only retrieved if they refer to a person.
2. It facilitates the retrieval of articles that mention semantic entities in different ways, but without the user having to enumerate all of these in their queries.
3. It allows users to specify how search terms should be related to each other to describe specific events.
4. It allows users to place restrictions on the discourse contexts of the events retrieved, to account for various event interpretations.

In the following sections, we look at various resources and methods we have used to facilitate the provision of such functionalities in our search system.

III. The New York Times Corpus

The NYT corpus consists of around 1.8 million articles, ranging in time from 1987 to 2007. Each article is annotated with various types of metadata, including basic general information, such as the author, date, day of the week and themed column in which the article appeared. This could allow the development of a search system that permits filtering based on article-level attributes, in a similar way to the Google News and Yahoo News search engines. However, the annotated metadata goes beyond this, to include semantically-oriented information about the content of the article, such as specific types of named entities mentioned within the article. Thus, the corpus can help to facilitate the development of

systems that fulfil the first criterion mentioned in the previous section, i.e., to allow users to place restrictions on the types of entities retrieved by a search. Furthermore, the named entities annotated in the corpus are *normalised*, based on a controlled vocabulary applied consistently across articles. So, for example, if there are two different articles that refer to the same person in different ways, e.g., "*Bill Clinton*" and "*President William Jefferson Clinton*", the normalised name "CLINTON, BILL" will be assigned as a metadata attribute in both cases. Such normalisation of entities is important to allow the development of systems allowing the retrieval of articles concerning a specific entity of interest to the user, regardless of the exact way in which it is mentioned in the text.

Although it provides support for recognising entities and their variants, the NYT corpus does not define how entities are linked together to describe *events*, nor does it consider the encoding of discourse-related phenomena. Therefore, it cannot aid in the development of a system that fulfils the third and fourth criteria specified in section II.

IV. EVENT-BASED SEARCHING

The main purpose of news articles is to report upon events that have occurred or are occurring in the world, e.g., attacks, arrests, murders, etc. Researchers searching the news are usually interested in finding information about specific events, e.g. celebrity marriages, or classes of events, as in the sample scenario involving Saddam Hussein. This means that it is sensible for search criteria in a news search system to be based around the specification of events, and to allow restrictions to be placed on the type(s) of events of interest, who is involved in these events and how, etc.

The bottleneck in developing event-based news search systems is that much richer metadata about the content of articles is required than the NYT corpus provides. To form the basis of such a system, the metadata for each article in the NYT corpus would need to be augmented to encode the details of all significant events in the article. Given the size of the corpus (i.e., 1.8 million articles), this is not a feasible manual task. Additionally, since thousands of news articles are published every day, it is important to be able to employ automated methods to extract such metadata, in order that search results include the most up-to-date information.

Accordingly, we use sophisticated text mining techniques, which are able to "understand" the content of articles and extract metadata corresponding to *structured* representations of events automatically for each article.

Consider sentence (1):

(1) *John Smith killed his wife in Texas in 1988.*

We use the term *event* to refer to a specific *textual* mention of a physical event. Thus, several textual mentions may map to a single physical event. The event in (1) is denoted by the verb *killed,* whilst other parts of the sentence provide details about the event *participants*, i.e., *John Smith* is the perpetrator, *his wife* is the victim, *Texas* is the location and *1988* is the time. Event-based search systems should allow users to specify restrictions on both the type of event to be retrieved and its participants, in an interactive manner.

To provide such functionality, our tools analyse the (syntactic) structure of the text to identify and characterise the event participants, e.g., to determine that the subject of the verb *killed* corresponds to the semantic agent of the event, etc.

In order to achieve the widest possible coverage of event recognition without the manual burden of writing rules, a well-established method is to *train* a system to recognise the appropriate information by applying machine learning techniques to a corpus of texts in which the events have been (manually) annotated (e.g., [5, 6]). The system learns generalised patterns and features of the text occurring in the context of different types of events and their participants, meaning that it is able to detect events in *unseen* texts, even if they take a slightly different form or are denoted using words other than those found in the training corpus.

A. The ACE Corpus

The ACE 2005 corpus [4] is a manually annotated corpus used as a basis for training event recognition systems. The types of text/events covered are relevant to news search, since a part of the corpus covers newswire (including NYT articles) and transcripts of broadcast news, whilst the remainder of the corpus mainly concerns discussions about news topics.

In total, there are 599 documents in the corpus, containing 5349 events. The corpus is annotated with named entities, *structured events* and coreference (linking together various different mentions of the same entity), the latter of which has been shown to be beneficial in improving the results of event recognition (e.g., [7, 8]).

Given the overlap of text types between the ACE 2005 corpus and the NYT corpus, we applied named entity/event extraction systems trained on the ACE 2005 corpus to the documents of the NYT corpus, in order to enrich them with additional semantic metadata corresponding to the events and named entities needed for event-based search.

According to the ACE event representation, the event in sentence (1) is represented as follows:

EVENT_TYPE: LIFE_DIE
TRIGGER: *killed*
AGENT: *John Smith*
VICTIM: *his wife*
PLACE: *Texas*
TIME: *1988*

For each event, a *type* is assigned, and a trigger (the word or phrase that best characterises the event) is identified. The assignment of event types allows a system to learn to recognise and categorise semantically similar events in text, regardless of the trigger word or phrase used. For example, a *LIFE_DIE* event could be denoted with nouns such as *death* and *execution* as well as verbs like *murder* and *pass away*. The

trained system can thus allow the user to specify only the *type* of the event that they are interested in, rather than having to worry about entering exact trigger words. The ACE corpus contains 33 different event types, falling under 8 different categories that are frequently reported in news stories, i.e. LIFE, MOVEMENT, TRANSACTION, BUSINESS, CONFLICT, CONTACT, PERSONNEL and JUSTICE.

The assignment of types or *semantic roles* (e.g. AGENT, VICTIM, etc.) to event participants additionally allows a system trained on the corpus to "learn" the various ways in which particular participant types can be expressed in text. This allows users to specify their search criteria as *semantic* restrictions on event types and participants, which abstract away from the various possible textual representations of the events to be retrieved. Returning to the search scenario of finding information about people killed by Saddam Hussein in Iraq, the following semantically-oriented search template would allow the researcher to carry out a more focussed search:

EVENT_TYPE: LIFE_DIE
AGENT: Saddam Hussein
PLACE: Iraq

The only articles retrieved by this template would be those mentioning deaths, and more specifically killings (according to the requirement for an AGENT of the LIFE_DIE event). The AGENT should be *Saddam Hussein*, and the PLACE should be *Iraq*. A greater or lesser number of restrictions can be included within the search template, according to users' needs.

V. Discourse Context of Events

In creating an interactive, semantic event-based system, it is desirable to go beyond identifying only event triggers and participants in text, and to also consider features of the event's discourse context, which can reveal both subtle and significant aspects about its interpretation. Consider the following sentences:

(2) *Hussein failed to carry out the planned executions in Iraq.*

(3) *According to an Iraqi government spokesperson, Saddam wants to carry out further murders in Iraq.*

(4) *Iraq's leader said that he will perform the executions on Thursday.*

(5) *We condemn all of Saddam Hussein's murders of his countrymen in Iraq.*

The events in sentences (2) to (5) would all match the search template specified in the previous section. They all mention instances of killings in which Saddam Hussein is the AGENT, and where the PLACE is Iraq. Despite this, none of these events is likely to be of interest to a researcher who wants to obtain information about specific past killings by Saddam. Only by considering the discourse context of the events can we discover their true interpretation. In sentence (2), the executions did not happen, while in sentence (3), Saddam is only speculating about carrying out the murders. In (4), the future tense is used, meaning that, at the time of the report, the executions had not taken place. Sentence (5) is more generic, in that it does not refer to any specific killings.

Although the researcher may not be interested in these events in the specific scenario mentioned, we should not exclude negative, non-definite or non-specific events from all search results. For example, another researcher may be interested in looking at cases where Saddam's attempts to kill people failed, or finding out how often his promises or threats to murder people actually turned into reality.

Therefore, our event-based search system allows users to filter their results according to the specific discourse contexts of events. This provides greater scope for users to eliminate irrelevant information from their results than if searches only involved event types and participants, whilst also allowing search results to be explored from several additional perspectives.

The ACE 2005 corpus assigns attributes to each event that encode certain aspects of their discourse context. The attributes assigned and their values are as follows:

- POLARITY — *Negative* if it is explicitly stated that the event did not take place, or *Positive* otherwise
- TENSE — *Past, Present, Future* or *Unspecified*. Assigned according to the time that the event took place with respect to the textual anchor time (i.e., the time of broadcast or publication). *Unspecified* is assigned if it is not clear when the event took place or if it has taken place.
- MODALITY — *Asserted* when the author or speaker makes reference to the event as though it were a real occurrence, and *Other* otherwise.
- SPECIFICITY — *Specific* if the event is understood as a singular occurrence at a particular place and time, or a finite set of such occurrences, or *Generic* otherwise.

The values of these attributes allow distinctions to be made between the different types of interpretations of the events in the above example sentences: the event in sentence (2) would be assigned the POLARITY value of *Negative*. Although the POLARITY value of the event in sentence (3) is *Positive*, the MODALITY value would be *Other,* since it is a speculated rather than a "real" event. The TENSE value for (3) would be *Present,* in contrast to sentence (4), whose TENSE value would be *Future,* but whose MODALITY value would be *Asserted,* since the event is stated as though it certainly will happen. In contrast to the events in sentences (2)–(4), whose GENERICITY value would be set to *Specific,* the event in sentence (5) would be assigned the *Generic* value.

VI. Refining Event-Based Search Through Enhanced Discourse-Level Annotation

Although annotated with several basic discourse-level attributes, the ACE 2005 corpus fails to identify various other

commonly occurring types of discourse phenomena. Some such phenomena can be observed in the example sentences in section V. In sentence (3), Saddam is not only speculating about carrying out further murders, he actually *wants* to carry them out, i.e., he has a positive subjectivity towards this possible event. In contrast, the reporter in sentence (5) expresses a negative opinion towards Saddam's killings. A further point about sentence (3) is that the information is stated as being provided by a specific source, i.e., the Iraqi government. Thus, this information could be highly biased.

The automatic recognition of the above types of information can be highly relevant in news articles. The expression of different sentiments and opinions in news articles has already been widely studied, e.g., [9, 10], because news stories are rarely reported in a neutral way [11]. Since a sentence may contain sentiments about multiple topics [12], the assignment of subjectivity values at the level of events can help to disentangle the sentiments that may be expressed towards different events in the sentence. The identification of information source is also very important, given that the percentage of sentences containing direct or indirect reported speech can be as high as 90% in some news articles [13]. Additionally, attribution to a particular source could either be done in a positive way, to bolster a claim made in the text already, or otherwise to distance the author from the attributed material, implicitly lowering its credibility [14].

The recognition of details about opinions and information source can be useful in studying potential contrasts and contradictions occurring in news articles (e.g., [15, 16]). Possible questions include: Which instances of events have contrasting opinions about them? Which information sources are responsible for these differing opinions? Is conflicting information provided by different information sources for some event? How reliable is the information provided by a particular source?

To allow training of systems that permit researchers to explore such questions more easily, we have enriched the ACE 2005 corpus with new information. This includes both the identification of additional types of discourse-level phenomena, and the refinement of existing phenomena in the corpus.

Prior to undertaking the enrichment, we studied a number of existing corpora of news articles annotated with discourse-level information. The corpus reported in [17, 18] concerns the annotation of statements occurring in news articles. Some of the attributes annotated are very similar to the information in the ACE 2005 corpus, e.g., the factuality of the statement (factual or abstract) and the time being referred to (past, present or future). However, the main focus of the work is on determining the level of certainty that can be ascribed to statements, amongst five possible levels. The MPQA corpus [19], meanwhile, is annotated for opinions, encoding information such as the intensity and polarity of the attitude expressed. The FactBank corpus [20] is perhaps the most comparable to the ACE corpus, in that information is annotated at the level of events, according to the observation that statements are not necessarily the ideal unit of text to which discourse-level information should be assigned: a statement may contain several events, each having a different interpretation. However, the events in FactBank are different to those in ACE, in that they do not identify event participants. FactBank is built on top of TimeBank [21], which identifies events and assigns temporal information (e.g., tense and aspect) and polarity (i.e., whether or not the stated event took place). FactBank enriches TimeBank by assigning factuality values to events, ranging from certainty that the event did take place to certainty that it did not, with various degrees in between.

Common to all three of the annotation efforts examined is the identification of the information source, i.e., the person, organisation, etc., that is the provider of the stated information. The annotation in [17] also makes a distinction between identifying the source as the author/writer, a direct participant or an uninvolved expert.

After reviewing the above resources, we refined the discourse-based annotations in the ACE 2005 corpus:

- The values of the POLARITY, TENSE and SPECIFICITY attributes remain the same.
- For MODALITY, we added a *Speculated* value, based on evidence from some of the corpora reviewed that various degrees of certainty or factuality of the information reported can be distinguished. Our decision to use a fairly simple distinction is based on the finding that only limited agreement between human annotators could be reached when using a more detailed scale [18].
- A further new value was also added for the MODALITY attribute, i.e., *Presupposed*, to account for events describing situations that are previously known or assumed within the discourse. According to the updated scheme, *Asserted* is only assigned when new events are introduced into the discourse. Consider the following: *A man shot two people last light. The shootings took place in New York*. According to the ACE corpus annotation scheme, two (textual) events would be annotated, one with the trigger *shot* and the other with the trigger *shooting*. The event in the first sentence would be assigned *MODALITY=Asserted,* since it introduces a new (physical) event into the discourse. However, the event in the second sentence refers to the same physical event, and so would be assigned *MODALITY=Presupposed*. Thus, the introduction of the *Presupposed* value makes it possible to isolate events that introduce new information into the discourse.
- Information about the SOURCE of the event is annotated in the updated corpus, based on the importance explained above. We identify the named entity corresponding to the source in the text (if the source is not the author) and also assign a label to categorise the source. In this respect, our approach takes inspiration from [17]. The category can

be *Author* (if the information is presented as coming from the current writer or speaker), *Involved* (if the information comes from somebody directly involved in the reported event) or *Third Party* (if the information comes from somebody who is not the author, and they are not directly involved in the event, e.g., a newspaper). The distinction between *Involved* and *Third Party* could assist, for example, in studying potential biases or the reliability of the information provided by different types of source.

- We have added a SUBJECTIVITY attribute, to allow information about opinions and subjectivity specified with respect to events to be explicitly identified and encoded. The information annotated is comparable to the attitude polarity in the MPQA corpus, in that it encodes whether *negative, positive* or *neutral* subjectivity/attitude is expressed towards an event. A *multi-valued* category is used to account for cases in which multiple types of subjectivity, both positive and negative, are specified in the context of a single event.

Based on the above changes to the annotation scheme, we reviewed and augmented the discourse-related attributes for all of the 5349 events in the ACE Corpus.

Existing discourse-related attributes were reviewed for a number of reasons:

- Since we updated the possible values of the MODALITY attribute, the values assigned to many events were subject to change (mainly from *Asserted* to *Presupposed*).
- This original annotation effort was based on rather sparse instructions, and we found that this had sometimes led to inconsistent annotations. To rectify this, we created an updated guideline document, with clearer instructions and definitions for both the existing and new discourse-related attributes, which was consulted during the re-annotation effort.
- We annotated cue words and phrases in the text that provide evidence for the assignment of specific values to the discourse attributes, e.g., speculative words like *may* or *probably* and subjective words like *hope* or *condemn*. This decision was made according to previous analyses of texts in the academic scientific domain, showing that various discourse-related information is very often conveyed using particular cue words and phrases (e.g., [22-24]). It has also been shown that the identification of such cues in an annotated training corpus can increase the accuracy of a system trained to recognise discourse-level information about events [25].

The annotation described above was mainly carried out by one of the authors, with computational linguistics expertise and several years' experience of working with annotation schemes. In order to verify the soundness of the annotation scheme and the quality of the annotations produced, a second annotator, also an author with computational linguistics expertise, annotated around one fifth of the corpus. We measured the agreement between the annotators on these 1000 events in terms of Cohen's Kappa [26], which is the standard way of measuring agreement between annotators. Averaging over the values of the six different discourse attributes, the agreement reached between the annotators is 0.76 Kappa. According to the interpretations of Kappa scores provided in [27], this score means that substantial agreement was reached between the two annotators, providing strong evidence that the new discourse-based annotation is of a high quality.

VII. Text Mining the New York Times Corpus

In this section, we provide a brief overview of our system (ISHER)[3] that puts into practice several of the semantic metadata extraction methods mentioned in the previous sections of this paper. Building on technology developed for the ASCOT clinical trials search system [28], ISHER provides users with an enhanced search experience that allows them to explore NYT articles from several semantically-motivated angles and to filter their search results according to various semantic criteria. The current version of ISHER is trained on the original version of the ACE 2005 corpus (i.e., prior to our enrichment of the corpus).

ISHER is Web-based, with an interface designed in such a way that users do not have to learn any new ways of formulating queries in order to take advantage of semantic information. Rather, they begin their search by entering keywords. The rich semantic metadata associated with each of the retrieved articles, obtained through the application of text mining techniques, is then used to allow the user to refine the results of this initial search in a multi-faceted way.

Retrieved news articles are presented to users along with semantic information that is arranged within several tabs. These tabs allow refining/filtering of the search results according to one or more of the following facets:

- Grouping of semantically similar documents. Articles retrieved by a search are automatically clustered into groups and assigned thematic labels. Users can refine their search by choosing clusters of interest to them. Groups of clusters can be visualised to show the semantic closeness of specific articles in a particular cluster to the thematic topic, and to see links between clusters.
- Metadata categories from the original NYT corpus.
- Named entities found through text mining analysis.
- Event types and participants found though text mining analysis.
- The values of the discourse-level attributes of POLARITY, MODALITY, GENERICITY and TENSE.

A screenshot of the system is shown in Figure 1. It illustrates a scenario in which the user is searching for mentions of CONFLICT events in which Saddam Hussein is identified as the ATTACKER, and whose MODALITY value is *Other* (i.e., the event does not necessarily refer to a definite event that has

[3] http://nactem.ac.uk/DID-ISHER/

happened or will happen). The top of the screen displays several event-based semantic restrictions that have been chosen, by selecting event features from the hierarchy displayed on the left of the screen, which include event and participant types, together with discourse-based attributes. Any number of restrictions may be chosen, in order to "drill down" to a set of articles of specific interest. The main part of the screen illustrates one of the 73 articles that have been retrieved according to specified restrictions, and in which the highlighted text corresponds to an event of interest to the user. The MODALITY value is *Other* because Saddam *threatened* to carry out attacks, but they had not yet been carried out at the time that the article was written. On the right hand side of the screen, various details about the selected event are displayed, including the values of the discourse-based attributes, the trigger and the identified participants (marked as *Role*).

VIII. Conclusions and Future Work

In this paper, we have described how the vast content of digital heritage data resources can be made more readily accessible and searchable, through the provision of semantic search capabilities. Our work has concentrated on enhancing access to information contained within a particular digital heritage archive, i.e., historical articles from the New York Times. We explained how semantic metadata in the annotated NYT corpus can help to improve more focussed access to documents and we motivated the need for the application of more sophisticated text mining methods to the archive, in order to facilitate event-based searching.

Furthermore, we discussed how the various discourse contexts in which events occur can significantly alter their interpretation, meaning that the automatic identification and classification of discourse phenomena provides useful additional criteria for searching and filtering events. We introduced the ACE 2005 corpus, and explained how its event and discourse-related annotations constitute a valuable source of data for training sophisticated event-based search systems.

We subsequently explored several ways in which the discourse-related information annotated in the ACE 2005 corpus misses important discourse phenomena that can be readily identified in the textual context of events, i.e., the information source of the event, subjectivity/opinions specified towards the event and whether there is any speculation surrounding the event. We explained how the automatic recognition of such phenomena can allow researchers to explore archives such as the NYT from various new angles, e.g., to identify and compare contrasting and conflicting reports about particular events and to study the reliability of information provided by different information sources. Based on the importance of recognising these types of information, we embarked upon a new annotation effort to enrich and enhance the previously available discourse information provided in the ACE 2005 corpus. The resulting corpus, in which the discourse-related annotations for all 5349 events in 599 documents have been reviewed and enhanced, constitutes a valuable data set for training sophisticated, discourse-aware event extraction systems. We finally presented ISHER as a concrete example of a semantically-based search system that provides access to the NYT archive, and is trained using the ACE 2005 corpus. The semantic filtering criteria provided in ISHER address most of the search system desiderata that we identified, including filtering according to classified named entities, structured events and basic discourse-based features.

Future work will include making use of the enhanced ACE 2005 corpus to develop a more sophisticated version of the ISHER system. Since ISHER is based on machine-learning and trained using the ACE 2005 corpus, it would be reasonably straightforward to enhance its functionality to take into account our updated model of the discourse interpretation of events. It would be required to train a new model on the updated corpus, and then to apply this to the NYT archive of articles. Users would then be provided with several types of additional discourse-based information and filtering criteria.

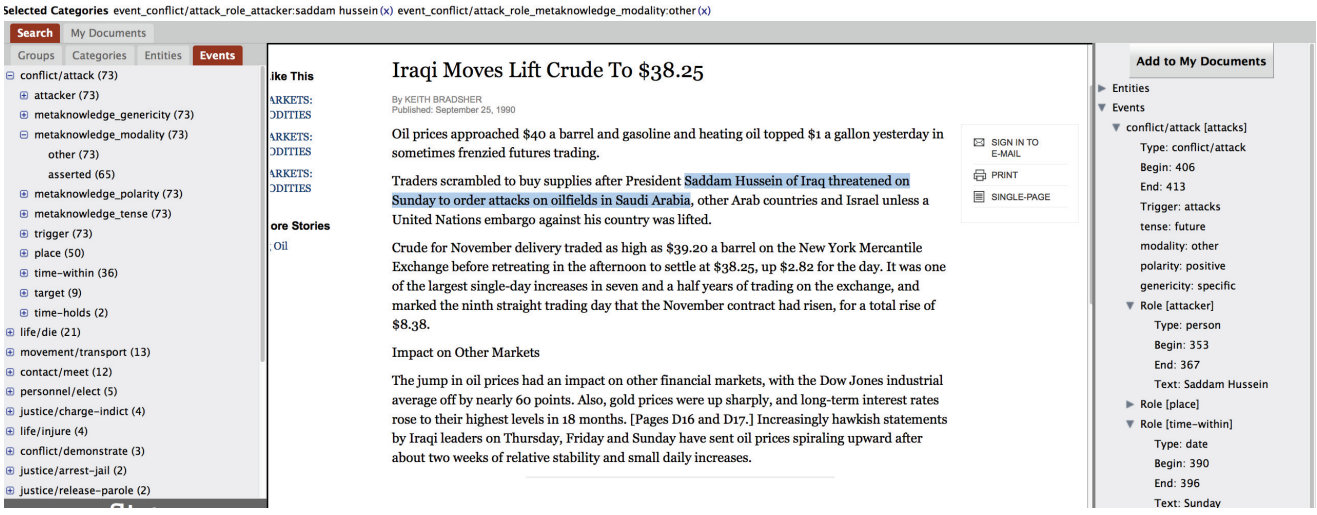

Fig. 1. Screenshot of the ISHER system showing event-based semantic restrictions and event features

REFERENCES

[1] E. Sandhaus, "The New York Times Annotated Corpus," Linguistic Data Consortium, Philadelphia, 2008.

[2] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A brief history," in Proceedings of the 16th International Conference on Computational Linguistics (COLING'96), pp. 466-471, 1996.

[3] S. Strassel, M. A. Przybocki, K. Peterson, Z. Song, and K. Maeda, "Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction," in Proceedings of the 6th Language Resources and Evaluation Conference, pp. 2706-2709, 2008.

[4] C. Walker, S. Strassel, J. Medero, and K. Maeda, "ACE 2005 multilingual training corpus," Linguistic Data Consortium, Philadelphia, 2006.

[5] S. Soderland, "Learning information extraction rules for semi-structured and free text," Machine Learning, vol. 34, pp. 233-272, 1999.

[6] M. E. Califf and R. J. Mooney, "Bottom-up relational learning of pattern matching rules for information extraction," Journal of Machine Learning Research, vol. 4, pp. 177-210, 2003.

[7] M. Miwa, P. Thompson, and S. Ananiadou, "Boosting automatic event extraction from the literature using domain adaptation and coreference resolution," Bioinformatics, vol. 28, pp. 1759-1765, pp. 19–26, 2012

[8] F. Celli and M. Poesio, "Improving relation extraction with anaphora resolution in Italian," in Proceedings of DAARC, 2011. Available at: http://clic.cimec.unitn.it/fabio/fc-mp11-daarc.pdf

[9] M. Bautin, L. Vijayarenu, and S. Skiena, "International sentiment analysis for news and blogs," in Proceedings of the International Conference on Weblogs and Social Media, 2008.

[10] A. Balahur, R. Steinberger, M. A. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," in Proceedings of the 7th Language Resources and Evaluation Conference, pp. 2216-2220, 2010

[11] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in Proceedings of the International Conference on Weblogs and Social Media, 2007. Available at: http://www.icwsm.org/papers/3--Godbole-Srinivasaiah-Skiena.pdf

[12] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," in Proceedings of the Third IEEE International Conference on Data Mining, pp. 427-434, 2003.

[13] S. Bergler, "Conveying attitude with reported speech," in Computing attitude and affect in text: Theory and applications, ed: Springer, pp. 11-22, 2006.

[14] P. Anick and S. Bergler, "Lexical structures for linguistic inference," in Lexical Semantics and Knowledge Representation, Springer, pp. 121-135, 1992.

[15] B. K. Bal, "Towards an analysis of opinions in news editorials: How positive was the year?," in Proceedings of the Eighth International Conference on Computational Semantics, pp. 260-263, 2009.

[16] P. Carvalho, L. Sarmento, J. Teixeira, and M. J. Silva, "Liars and saviors in a sentiment annotated corpus of comments to political debates," in Proceedings of ACL (Short Papers), pp. 564-568, 2011.

[17] V. Rubin, E. Liddy, and N. Kando, "Certainty identification in texts: Categorization model and manual tagging results," Computing attitude and affect in text: Theory and applications, J.G. Shanahan, Y. Qu and J. Wiebe, Eds. pp. 61-76, 2006.

[18] V. L. Rubin, "Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts," Information Processing & Management, vol. 46, pp. 533-540, 2010.

[19] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," Language Resources and Evaluation, vol. 39, pp. 165-210, 2005.

[20] R. Sauri and J. Pustejovsky, "FactBank: A corpus annotated with event factuality," Language Resources and Evaluation, vol. 43, pp. 227-268, 2009.

[21] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, and L. Ferro, "The TimeBank corpus," in Proceedings of Corpus Linguistics, pp. 647-656, 2003.

[22] K. Hyland, "Talking to the academy: Forms of hedging in science research articles," Written Communication, vol. 13, pp. 251-281, 1996.

[23] V. Rizomilioti, "Exploring epistemic modality in academic discourse using Corpora," in Information Technology in Languages for Specific Purposes, E. Arnó Macià, A. Soler Cervera, and C. Rueda Ramos, Eds., New York: Springer, pp. 53-71, 2006.

[24] P. Thompson, G. Venturi, J. McNaught, S. Montemagni, and S. Ananiadou, "Categorising modality in biomedical texts," in Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining, Marrakech, Morocco, pp. 27-34, 2008.

[25] M. Miwa, P. Thompson, J. McNaught, D. B. Kell, and S. Ananiadou, "Extracting semantically enriched events from biomedical literature," BMC Bioinformatics, vol. 13, p. 108, 2012.

[26] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, pp. 37-46, 1960.

[27] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," Family Medicine, vol. 37, pp. 360-363, 2005.

[28] I. Korkontzelos, T. Mu and S. Ananiadou, "ASCOT: a text mining-based web-service for efficient search and assisted creation of clinical trials," BMC Medical Informatics and Decision Making, 12(Suppl 1), S3, 2012.