Human Language Technologies in Media Consumption: The Case of SAM

David Tomás and Yoan Gutiérrez
University of Alicante, Carretera San Vicente s/n, San Vicente del Raspeig, 03690, Spain
{dtomas,ygutierrez}@dlsi.ua.es

***Abstract*—This paper describes the role of human language technologies (HLT) in the framework of SAM, an EU-funded research project that focuses on developing an advanced digital media delivery platform based on second screen interaction and content syndication within a social media context, providing open and standardised ways of characterising, discovering and syndicating digital assets. HLT technologies are employed in SAM to analyse user comments in the context of media consumption, offering both a better experience for the end users and improved business intelligence reports for content providers. The tasks covered by these technologies in SAM include ontology development, entity linking, sentiment analysis, emotion detection, ontology matching, and text summarisation. This paper describes the approaches followed and the benefits of these technologies in the framework of social media and content consumption.**

***Index Terms*—ontology development, entity linking, sentiment analysis, emotion detection, ontology matching, text summarisation, social media**

I. Introduction

The current generation of Internet devices has changed how users are interacting with media, from passive and unidirectional users to proactive and interactive. Users can comment or rate a TV show and search for related information regarding characters, facts or personalities. This phenomenon is known as *second screen*. In today's second screen ecosystem there are no true standards, protocols or commonly used frameworks through which users can discover and access information related to consumed contents. Users have to actively perform searches using web search engines such as Google to participate in their favourite TV shows.

SAM[1] (Socialising Around Media) is an EU-funded project focused on developing an advanced digital media delivery platform for second screen and content syndication within a social media context, providing open and standardised ways of characterising, discovering and syndicating digital assets (e.g. films, songs, books, metadata, etc.). The project, currently under development, started in September 2013 and will finish in October 2016, with a duration of 37 months.

The potential customers of SAM are both business stakeholders (such as media broadcasters, content asset providers, software companies and digital marketing agencies) and end users. For the former, this platform provides a number of benefits, including dynamic social and media content syndication, managing of online reputation, listening to customers, tracking real time statistics or monitoring media related social content through second screen. For the latter, SAM offers a complete solution for people consuming media and TV programs. The platform integrates context aware information and complex social functionalities providing contextual information about actual and current interests of the user. These features create an augmented experience in which end users can discover new information about the subject, talking and sharing their experiences with other users that are also interested in the same topic.

Some of the core functionalities of SAM are based on human language technologies (HLT), i.e., the application oriented use of natural language processing (NLP) to enable computers to derive meaning from human or natural language input. The NLP tasks addressed in SAM include ontology development, entity linking, sentiment analysis, emotion detection, ontology matching, and text summarisation. This paper describes these tasks, the approaches developed in SAM, and its benefits in the context of social media and content consumption.

The remainder of this article is organised as follows: next section describes the three core technologies supporting the SAM platform; Section III describes the NLP challenges addressed in SAM and the decisions taken; Section IV provides information about existing platforms in the area of business of SAM; finally, Section V provides conclusions and describes remaining work.

II. The Three Pillars of SAM

The SAM platform has been designed around three pillars that highlight the main research and business directions of this project: content syndication, second screen and social media. By combining these pillars together, SAM implements the distribution of media assets to the end users through their devices, including linked content and related information to enrich the user experience.

a) Content Syndication: Technologies like Really Simple Syndication (RSS) have their place in the syndication world, but when an organisation needs to push more enriched and target-adapted information to partner websites or social networking sites (from product details to full microsites with rich media), approaches such as RSS are not powerful enough. Content syndication [1] solutions today have evolved beyond RSS, allowing vendors, distributors and publishers to issue, control and track rich content experiences on third-party websites. In SAM, content syndication techniques will allow content providers to prepare their digital assets and associate

[1]http://www.socialisingaroundmedia.com.

them to specific media and usage context, offering mechanisms for these enriched assets to be delivered in the expected format and to be consumed by the users in a specific context.

b) Second Screen: This concept refers to any electronic device (broadly a mobile device, such as a tablet or smartphone) that allows users to retrieve additional information about the content they are watching on the first screen (usually a TV set). The SAM platform includes a multi-device representation layer that provides syndicated information in the appropriate format to be consumed by different types of devices. This generic approach is used in order to access the asset-related syndicated information while it is being consumed, commented, or interacted with by the users, creating a second screen experience.

c) Social Media: These technologies are changing the way in which users interact and communicate with each other, expressing their feelings, opinions and thoughts about almost anything, including products, personalities, and TV shows. In social media, users not only share comments or articles, they also exchange different types of digital assets such as videos, photographs and documents. In recent years, the user activity in social networks has significantly increased, making it into a key area of interest for media business and advertisers. Decision makers try to find ways in which commercial products can make profitable use of applications such as YouTube, Facebook and Twitter. In SAM, social interaction around digital media items provides the context in which the syndicated content is consumed. SAM incorporates complex context extraction mechanisms that rely on NLP technologies for the creation of dynamic social communities based on users' actions and their context (e.g. assets consumed, demographic profile and preferences).

III. NLP Technologies in SAM

As mentioned before, some of the core functionalities of the SAM platform are supported by different NLP technologies. This section describes the main tasks that will be faced in this project by means of these technologies and its benefits in the context of social media and content consumption. Figure 1 shows the high-level architecture of the SAM platform. The Semantic Services module shown in this figure contains the subcomponents in charge of providing the NLP functionalities described in this section: Sentiment Analysis (performs sentiment analysis and emotion detection as described in Section III-C and Section III-D, respectively), Text Summarisation (performs the summarisation of texts described in Section III-F), and Data Characterisation (offers entity linking and ontology matching as described in Section III-B and Section III-E, respectively). The Asset Profiler and the Asset Discovery rely on the NLP technologies provided by the aforementioned subcomponents. Finally, the Cloud Storage component stores the ontology definition and instances described in Section III-A.

A. Ontology Development

An ontology is a formal specification of a shared conceptualization, an abstract and simplified view of the world that we wish to represent for some purpose. Every knowledge-based system is committed to some conceptualization, explicitly or implicitly [2]. An ontology describes the objects, concepts, and other entities that are assumed to exist in some area of interest (i.e. a particular domain) and the relationships that hold among them [3]. The use of ontologies for knowledge representation allows sharing common understanding of the structure of information among people or software agents, enabling reuse of domain knowledge, making domain assumptions explicit, separating domain knowledge from the operational knowledge and analysing it [4].

Although ontologies are not only and exclusively an NLP resource, they are a key element in many language technologies that work at semantic level. In SAM, an ontology have been developed to model all the information related to the assets introduced in the platform and their owners. The ontology includes concepts such as `Person` and `Film`, attributes such as `familyName` and `subtitleLanguage`, and relations that identify, for instance, that a `Person` is a contributor in a `Film`. All the asset information is stored as instances of the SAM ontology. Sesame[2] is used as the framework for processing and handling these instances.

The SAM ontology reuses, were possible, concepts defined in the Europeana Data Model (EDM)[3] and Schema.org[4] in order to conform to existing standards. The EDM aims at being an integration medium for collecting, connecting and enriching the descriptions supplied by content providers in Europeana,[5] the Internet portal that offers an interface to millions of cultural heritage objects (e.g. books, paintings, films, and archival records) that have been digitised throughout Europe. Schema.org is a collection of shared vocabularies that webmasters can use to mark-up HTML pages to be understood by the popular search engines, such as Bing, Google or Yahoo!. This mark-up enables search engines to understand the information on web pages and provide richer search results in order to make it easier for users to find relevant information on the web. Besides reusing existing concepts from these two schemas, in SAM new ones were proposed to fulfil the project's requirements and give a proper coverage to all the possible input information distributed into the platform, from films synopsis to social networks links.

B. Entity Linking

Entity linking is the task of matching a textual entity mention to a knowledge base, such as a Wikipedia page, that is a canonical entry for that entity [5]. For instance, given a mention in a text to "Al Pacino", the goal of this task is to determine that it refers to the entity described in this specific entry in Wikipedia: `http://es.wikipedia.org/wiki/Al_Pacino`. This task is more challenging than traditional named entity recognition (NER), where the goal is to determine the occurrences of names in text and their classification. In the previous example, a NER system would

[2]http://rdf4j.org/.
[3]http://pro.europeana.eu/page/edm-documentation/.
[4]https://schema.org/.
[5]http://www.europeana.eu/portal/.

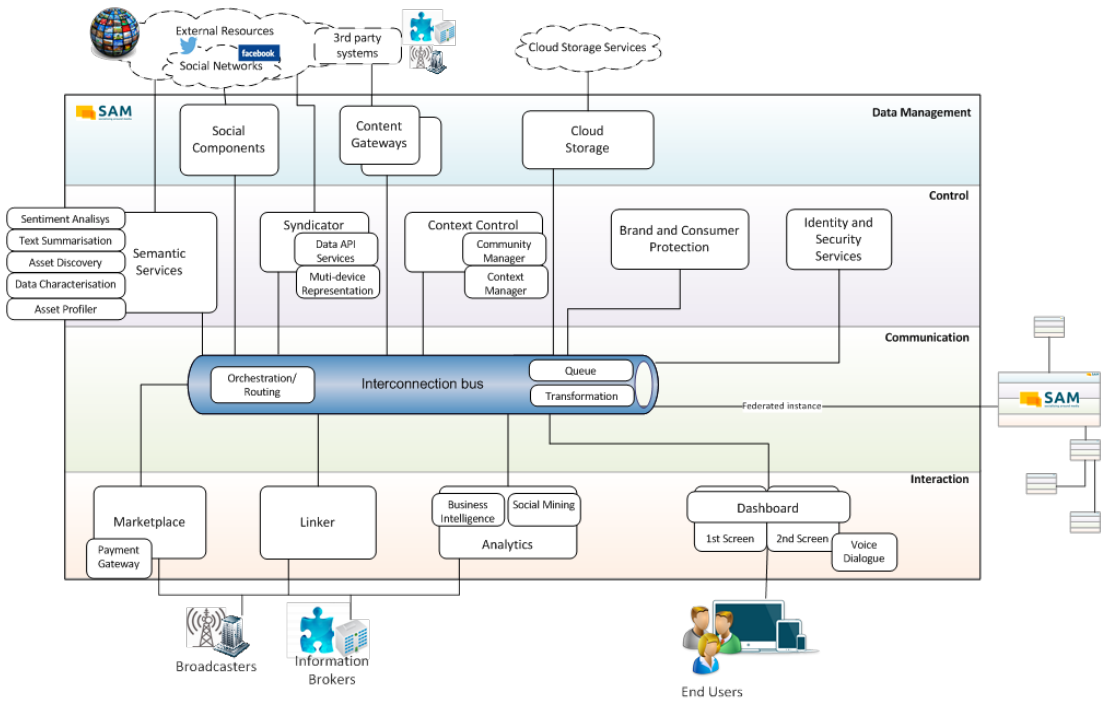

Fig. 1. The SAM platform high-level architecture. The NLP technologies developed in the project are provided by the Semantic Services subcomponents in the left side of this figure: Sentiment Analysis, Text Summarisation, Asset Discovery, Data Characterisation, and Asset Profiler.

determine that "Al Pacino" is a `person` or that "Los Angeles" is a `location` [6]. Entity linking requires a NER system, but this process must be complemented by a following disambiguation phase where this person or location is linked to an unambiguous entity stored in a knowledge base.

In SAM, entity linking technologies are applied on two different knowledge bases. First, the system identifies and links mentions in text to related Wikipedia pages. Secondly, it also identifies references to instances contained in the SAM ontology (e.g. books, songs, films, actors, etc. stored in the platform).

As mentioned in the introduction of this article, the potential customers of SAM are both business stakeholders and end users. For the former, entity linking provides a number of benefits in SAM, including the enrichment of their contents by linking them to additional internal (media assets stored in the platform) and external (Wikipedia pages) sources of information. The system also identifies comments on specific assets made by end users of the platform for business intelligence purposes (e.g. analysis of customer opinions about a specific content). Regarding the benefits for end users, entity linking provides an augmented experience in which users can discover new information about an asset, creating richer experiences around the original contents. For instance, a user watching the film "Casino Royale" in the SAM platform, and thanks to the entity linking module, would get additional information related to actors "Daniel Craig" and "Mads Mikkelsen" from Wikipedia, and also to other related assets in our platform based on the linking to our ontology, such as books created by "Ian Fleming", the writer of the series of spy novels.

The following paragraphs briefly describe the underlying technologies to entity linking developed in SAM.

1) Linking to Wikipedia: The approach to entity linking on Wikipedia is based on OpenNLP,[6] DBpedia,[7] and Lookup.[8] OpenNLP is employed to identify noun phrases and named entities that comprise the set of candidate terms to be linked to Wikipedia. Instead of working directly on Wikipedia, we work with DBpedia, its structured version, which facilitates querying and further processing by electronic means.

One of the main challenges in entity linking is that of name variations: an entity often has multiple forms, including shortened forms (Leonardo DiCaprio / Leo DiCaprio), aliases (Dwayne Johnson / The Rock), alternate spellings (Osama / Ussamah) and abbreviations (British Broadcasting Corporation / BBC). This problem is addressed in SAM by means of Lookup. Given a set of keywords, this tool retrieves related resources, either because the label of the resource or an anchor text frequently used in Wikipedia to refer to that resource matches the query. This functionality avoids the need for a perfect matching between the query and the resource, overcoming the problem of shortened forms, aliases and abbreviations.

Another relevant challenge is entity ambiguity: a single mention to an entity can match multiple knowledge base entries, since many entities tend to be polysemous. For instance, "Francis Bacon" can refer both to the English philosopher and to the Irish artist. In SAM, this problem is faced using three different strategies:

- Number of inlinks. In the Web, an inlink to a webpage X is a URL of another webpage which contains a link

[6]https://opennlp.apache.org/.
[7]http://wiki.dbpedia.org/.
[8]http://wiki.dbpedia.org/projects/dbpedia-lookup/.

pointing to X. This disambiguation method ranks the candidate DBpedia entities taking into account the number of inlinks pointing at those pages from other DBpedia pages. For instance, a mention to "Francis Bacon" in a text would determine that it refers to the philosopher and not to the artist, since the page in DBpedia for the former has 825 inlinks, whereas the entry for the later has 244.

- Context distance. This method compares the text surrounding the mention to a candidate entity (at sentence level) with the description contained in DBpedia, using the classical Lesk algorithm for word sense disambiguation [7]. Continuing with the previous example, if the sentence containing a mention to "Francis Bacon" is more similar to the description provided by the DBpedia entry of the English philosopher, than that of the Irish artist, the system would choose the philosopher as the best candidate for linking. The similarity between terms is computed applying the Levenshtein distance, defined as the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other [8].
- Hybrid approach. This method combines the two previous ones. The number of inlinks and context distance values are computed and normalised between 0 and 1. The resulting value is obtained by computing the average between these two values.

The following list summarises the steps carried out by the system to perform the linking:

1) Given an input text, Open NLP identifies noun phrases or named entities as candidates for the linking process. Currently we are experimenting with both possibilities: to extract noun phrases, which gives a wider set of candidates (better recall), and to extract named entities, which offers a more restricted set (better precision).
2) Lookup takes the previous candidates (noun phrases or named entities) as an input and retrieves a set of entries from DBpedia for each one. These entries can be filtered taking into account the ontology classes[9] and categories[10] existing in DBpedia entries.
3) After that, the disambiguation algorithm is applied to rank the resulting DBpedia entries for each candidate entity.

Experiments already done in the platform revealed that using named entities as input and inlinks for disambiguation obtained the best precision (67.66%) and F1 score[11] (70.34%), whereas using noun phrases and inlinks provided the best recall (86.30%) [9].

2) Linking to the SAM ontology: The approach to entity linking on the SAM ontology uses OpenNLP and Lucene[12] as its core tools. OpenNLP is used again to identify noun phrases. Lucene is a high-performance, full-featured text search engine library. In SAM it is employed to index and retrieve instances of our ontology related to a specific query (taking the form of a string of keywords).

The process of entity linking described here requires an initial offline process, where all the instances of our ontology are indexed by Lucene. These instances consist on information about media assets, e.g. metadata about a book including its title, writer, publisher, synopsis, number of pages, etc. The content of each asset is indexed as a separate document by Lucene. After this offline process, the following steps are required to identify entities in incoming texts and link them to our ontology:

1) Open NLP acts on the input text to identify noun phrases as candidates for the linking process. Unlike the previous approach, we did not retrieve named entities as an alternative method. This strategy would limit too much the number of candidate entities, reducing the recall of the system, especially taking into account that our ontology is under development and the number of instances contained is far from Wikipedia.
2) Noun phrases are sent as queries to Lucene and the more relevant documents (media assets) are retrieved for each request. Lucene is configured to perform fuzzy search queries, using a similarity measure based on the Damerau-Levenshtein algorithm [10]. In this way the problem of name variations is reduced.
3) Finally, the disambiguation algorithm based on context analysis is applied to re-rank the results retrieved by Lucene and provided the final list of possible instances of the SAM ontology associated to the noun phrases detected in the original text. Since currently there is no information on inlinks in the SAM ontology, the other two disambiguation methods mentioned above cannot be applied in this approach.

C. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is the task dealing with the automatic detection and classification of opinions expressed in text written in natural language [11]. This task combines techniques from different research areas, including importantly NLP and text mining. The aim of this technology is twofold: first, it tries to detect the polarity of the opinions made by people, identifying whether these opinions are positive, negative or neutral regarding a product, a brand, a service, an event, etc.; second, it tries to identify the intensity of these opinions, i.e., whether these positive or negative feelings are strong or weak. There is a fine-grained variant of opinion mining known as aspect-based opinion mining. The goal in this case is to determine opinions expressed on specific aspects of an entity, such as the battery or the screen of a mobile phone.

The automatic extraction of subjective information is a technological challenge, but also an opportunity for many kinds of businesses due to the possibilities that these technologies offer to companies. There are some relevant applications of opinion mining in this area: direct application to marketing, since it provides information on persons' opinions concerning a brand, a product, or even a specific characteristic, allowing businesses

[9]http://mappings.dbpedia.org/server/ontology/classes/.
[10]https://en.wikipedia.org/wiki/Help:Category.
[11]The harmonic mean of precision and recall.
[12]https://lucene.apache.org/.

to assess their strengths and weakness as perceived by potential customers; creation of user profiles, by analysing comments from a user regarding entities such as products or services in order to determine user preferences for and attitudes towards these entities; and facilitation of reputation management by monitoring and managing opinions about entities to identify opinion leaders, hot topics and influential users.

SAM applies sentiment analysis on social media to extract valuable information from users' comments, likes and assets consumed in the platform. The benefit provided to end users by this technology is the discovery of user preferences and attitudes while using the SAM platform. This information can be used to provide better asset recommendations and support the creation of dynamic social communities based on user preferences and feelings. With regards to business intelligence, sentiment analysis allows content providers in SAM, such as media producers, publishers, and broadcasters, to discover opinions, tendencies, or specific reactions in specific film sequences to further support decision-making.

The approach to sentiment analysis in SAM is based on supervised machine learning and text categorisation techniques combined with a ranking skip-gram system. The first step carried out by the system consists on creating a sentiment lexicon by adding lexical patterns: n-grams and skip-grams. An n-gram is a sequence of n consecutive words found in text. For instance, in the sentence "The movie is very predictable", "very predictable" is a 2-gram, and "The movie is very" is a 4-gram. A skip-gram is a generalisation of n-grams where words might be skipped and do not need to be consecutive. In the previous example, "is predictable" is an example of 1-skip-2-gram (a 2-gram with 1 skipped word).

These lexical patterns are extracted based on semantic and sentiment evidences obtained from text with a priori annotated sentiment polarity information (i.e., we know whether these texts express a positive, negative or neutral opinion). In order to determine the intensity expressed by these patterns for each polarity, we make use of our own scoring algorithm [12]. This algorithm takes into account the following features for each n-gram: the number of skipped terms (in case we use skip-grams), the number of occurrences of the n-gram in the corpus, and the proportion of occurrences for a specific polarity (for instance, how many times the word "smile" appears in positive texts with respect to the total number of occurrences of this word in the corpus).

The lexicon created in this way is then used to extract the features to feed a machine learning algorithm, SVM [13] in our case, which learns to classify the polarity of new incoming text.

This approach has been already evaluated on Twitter datasets used on international competitions, such as Task 2 at SemEval 2013[13] and Task 1 at TASS 2013[14], where the goal was to identify whether a tweet was positive, negative or neutral regarding a topic. These datasets involve over 20.000 tweets on a range of topics, including entities (e.g. "Steve Jobs"), products (e.g. "Kindle"), and events (e.g. "NHL playoffs"). Our system achieved 65.5% F1 score in the SemEval dataset, with a recall of 62.4% and precision of 62.7%. Regarding the TASS dataset, the best F1 score obtained was 56.9% with a precision of 61.7% and a recall of 52.8%.

An additional experiment was carried out in a context closer to the SAM business area using the Movie Review dataset[15], which includes a large amount of user reviews classified regarding its polarity. The system obtained very promising results on this dataset, achieving 88.6% F1 score, with a precision and recall of 88.6%. The difference with respect to the previous two experiments can be justified by the limited context information provided by tweets (a maximum of 140 characters) and the informal nature of their writing style.

D. Emotion Detection

Another functionality provided by NLP technologies in SAM is the detection of emotions in text. The concepts of emotion and opinion are not equivalent. Emotions refer to our subjective feelings and thoughts. Many opinion sentences express no emotions (e.g. "The voice of this phone is clear"), and many emotion sentences give no opinion (e.g. "I am so surprised to see you."). According to [14], people have six primary emotions: love, joy, surprise, anger, sadness and fear. These basic emotions can be subdivided into many secondary and tertiary ones.

Similarly to sentiment analysis, emotion detection can be treated as a classification task. While sentiment analysis is mainly concerned with specifying positive or negative opinions, emotion detection is concerned with detecting different feelings in text. As in sentiment analysis, emotion detection can be implemented using a machine learning based approach although the lexicon based approach is used more widely [15].

Lexical resources such as WordNet Affect [16] are really useful for detecting emotions in text generated by humans. This resource contains a set of affective concepts correlated with affective words. For example, the adjective "cheerful" is semantically linked to the name "cheerfulness", to the verb "cheer up" and to the adverb "cheerfully", being all of them related to the emotion class "joy". Using this resource in combination with machine learning algorithms can provide a relay effective way of detecting emotions in texts generated by humans.

In SAM, we approach this task by building an inference core based on Lucene, in which each emotion is indexed by a list of keywords and a large list of examples of their use in context. In order to populate both lists, the semantic relations between WordNet Affects labels and WordNet synsets (groups of terms that are considered semantically equivalent) [17] were used. To this end, all the data stored in WordNet for every synset was employed to populate the list of keywords (extracting the lemmas) and the list of examples (extracting the glosses, i.e., an explanation of the synset). Moreover, also information extracted from its hyponyms[16] were included for every synset. In order to identify an emotion in an input text, this text is used

[13]https://www.cs.york.ac.uk/semeval-2013/task2/.

[14]http://www.sngularmeaning.team/TASS2013/tasks.php.

[15]https://www.cs.cornell.edu/people/pabo/movie-review-data/.

[16]Y is a hyponym of X if every Y is a (kind of) X (dog is a hyponym of canine).

to query Lucene and retrieve the most probable emotion based on the similarity between this text and the lists of keywords and examples indexed for this emotion.

E. Ontology Matching

As mentioned in Section III-A, the information in the SAM platform is stored as instances of the SAM ontology. In order to facilitate content providers the process of importing their information into the SAM platform, it is necessary to match the original structure in their data to the SAM ontology definition.

SAM provides NLP functionalities to determine the correspondences between the input data and the SAM ontology concepts, finding classes of data that are semantically equivalent in order to facilitate their alignment. In this way, external data is integrated into the SAM platform and can be linked to already existing assets and other external resources, as described in Section III-B. For instance, an attribute `born` of type `date` in the incoming data would be matched to the attribute `dateOfBirth` in the SAM ontology, and an attribute `nof_players` of type `integer` would be matched to `numberOfPlayers` in SAM.

This matching is currently based on semantic structural algorithms [18]. These algorithms rely on the Levenshtein distance (already described in Section III-B1) in order to establish a measure to get the best candidate concepts for the alignment. Furthermore, a search is performer for the most similar structures by exploring adjacent semantic relations, calculating the most overlapping structures. As a result, a confidence score of the most similar matching terms from both data models is obtained.

F. Text Summarisation

Text summarisation is the process of automatically reducing a text document in order to create a summary that retains the most important points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. The main idea of summarisation is to find a representative subset of the data, which contains the information of the entire set. Document summarisation, tries to automatically create a representative summary or abstract of the entire document by finding the most informative sentences.

Generally, there are two approaches to automatic summarization: extraction and abstraction. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is as closer as possible to what a human might generate. Such a summary might contain words not explicitly present in the original text. Research into abstractive methods is an increasingly important and active research area. However, due to complexity constraints, research to date has focused primarily on extractive methods.

Summarisation technologies are used nowadays in a large number of business areas, and there is a considerable number of applications in which automatic summarisation provides competitive advantages: automatic summarisation of scientific papers for researchers, generation of newsletters for groups of experts, summarisation of search engine results about a specific topic, and generation of opinion reports about a product or service. It is in this field that SAM benefits from summarisation. The goal in SAM is to provide the most relevant comments made by users about a content inside the platform to the owner of this content. This is part of the advanced business intelligent report generated in the platform for content providers, which can in this way get a summary of what users say about their products.

In SAM, an extractive method approach to summarisation is followed: large amounts of user generated content (comments submitted to the platform while consuming digital contents) is summarised to keep the most representative ones. For instance, an input such as "Casino Royale is a great movie. And a PHENOMENAL Bond movie. I don't care what any of you have to say! #DanielCraig #007 #Bond #Poker", would be reduced to "Casino Royale is a great movie". The summarisation component in SAM includes a parameter to specify the compression ratio, i.e. the percentage of reduction that we want to keep from to the original text. For instance, a value of 20% will include only this amount of comments for the business intelligence report, discarding the remainder 80%.

IV. Related Work

This section describes existing platforms and companies in the main business areas of SAM, i.e., second screen and social TV.

Contentwise[17] is a content personalisation system offering recommendations and predictive browsing for broadcast companies, using third-party data to create an information-rich experience for users. Contentwise pulls data from services such as IMDB, Wikipedia, and social media websites like Facebook and Twitter. The platform provides user activity metrics, financial metrics, engagement metrics and recommendation effectiveness metrics to content providers.

Horizon Go[18] is a Social TV platform offering live channels and video on demand. The platform provides information about the most popular movies and TV shows, including information about what the friends of the user and other visitors like to watch. Horizon Go also suggests shows and movies that the users might like to watch based on their past viewing history. The platform integrates applications accessible through the TV set and other screens (e.g. smartphones and tablets) including YouTube, Facebook, Wikitrivia, latest news, weather forecast and traffic information.

Tivin[19] is an interactive TV platform that automatically synchronise the broadcast transmission on the air through the TV audio offering a second screen experience. The platform allows the interaction with social media (e.g. Facebook and Twitter) and the real-time measurement of television audience of a particular program, commented on and discussed by the

[17] http://www.contentwise.tv/.
[18] https://www.horizon.tv/en_ie/.
[19] http://www.tivin.it/.

viewers. Tivin offers features such as the approval rating of a TV program in real time, related and additional content, language subtitles, merchandising and polls.

Tellybug[20] is a platform that makes lists of recommendations based on which shows are trending in Twitter. Users can sign in with their Twitter credentials to see what their contacts have been tweeting about specific shows, while also marking shows as liked and disliked. Their algorithm to filter tweets by show is based on a mixture of hashtags, show titles and matching of other words and phrases.

Despite the functionalities offered by the existing social TV and second screen platforms, SAM remains unique in that it supports content providers, broadcasters and end users with advanced capabilities that rely in HLT:

- Asset management and characterisation based on ontology matching and entity linking.
- Sentiment analysis, emotion detection, and text summarisation applied to Business Intelligence.
- Creation of a linked content ecosystem thanks to entity linking functionalities.
- Dynamic content recommendations based on asset mentions in text.
- interoperability with cultural organisations by using an ontology based on EDM and Schema.org.

V. Conclusions and Future Work

This paper presented an overview of the NLP technologies developed in the framework of the SAM project. The goal of SAM is to provide a social media delivery platform based on second screen and content syndication. In this context, NLP technologies provide core functionalities for both end users and content providers, helping to store and query information in the platform, enrich data, improve user experience, and analyse opinions expressed by users regarding the contents consumed.

The project started on September 2013 and will run for 37 months, finishing by the end of October 2016. SAM is currently in its second year, and at this stage there are prototypes available for all the NLP functionalities required by the platform. The most relevant of these functionalities were described in this paper: ontology development, entity linking, sentiment analysis, emotion detection, ontology matching, and text summarisation. The goal for the last year of the project is to improve and evaluate the existing prototypes, both intrinsically (considering the NLP components in isolation to characterize their performance) and extrinsically (considering the components as part of the whole SAM platform).

Acknowledgment

This work has been partially funded by the European Commission under the 7th Framework Programme for Research and Technological Development through the SAM (FP7-611312) project.

[20]http://www.tellybug.com/.

References

[1] K. Werbach, "Syndication–the emerging model for business in the internet era," *Harvard business review*, vol. 78, no. 3, pp. 84–93, 2000.
[2] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquis.*, vol. 5, no. 2, pp. 199–220, 1993.
[3] M. R. Genesereth and N. J. Nilsson, *Logical Foundations of Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1987.
[4] N. F. Noy and D. L. Mcguinness, "Ontology development 101: A guide to creating your first ontology," Stanford University, Tech. Rep., 2001.
[5] D. Rao, P. McNamee, and M. Dredze, "Entity linking: Finding extracted entities in a knowledge base," in *Multi-source, Multilingual Information Extraction and Summarization*, ser. Theory and Applications of Natural Language Processing, T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, Eds. Springer Berlin Heidelberg, 2013, pp. 93–115.
[6] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.
[7] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proceedings of the 5th Annual International Conference on Systems Documentation*, ser. SIGDOC '86. New York, NY, USA: ACM, 1986, pp. 24–26.
[8] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.
[9] D. Tomás, Y. Gutiérrez, and F. Agulló, "Entity linking in media content and user comments: Connecting data to wikipedia and other knowledge bases," in *Proceedings of the Twenty Fifth eChallenges Conference*. IIMC International Information Management Corporation, 2015, in press.
[10] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, no. 3, pp. 171–176, 1964.
[11] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.
[12] J. Fernández, Y. Gutiérrez, J. M. Gómez, and P. Martinez-Barco, "Gplsi: Supervised sentiment analysis in twitter using skipgrams," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, 2014, pp. 294–299.
[13] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.
[14] W. G. Parrott, *Emotions in social psychology: Essential readings*. Psychology Press, 2001.
[15] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.
[16] C. Strapparava and A. Valitutti, "Wordnet affect: an affective extension of wordnet," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*. Lisbon, Portugal: European Language Resources Association, 2004.
[17] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[18] P. Shvaiko and J. Euzenat, "Ontology matching: State of the art and future challenges," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 1, pp. 158–176, 2013.