Automated Semantic Tagging of Speech Audio

Yves Raimond, Chris Lowis,
Roderick Hodgson
BBC R&D
London, United Kingdom
{yves.raimond,chris.lowis,roderick.hodgson}@bbc.co.uk

Jonathan Tweed
MetaBroadcast
London, United Kingdom
jonathan@metabroadcast.com

ABSTRACT

The BBC is currently tagging programmes manually, using DBpedia as a source of tag identifiers, and a list of suggested tags extracted from the programme synopsis. These tags are then used to help navigation and topic-based search of programmes on the BBC website. However, given the very large number of programmes available in the archive, most of them having very little metadata attached to them, we need a way to automatically assign tags to programmes. We describe a framework to do so, using speech recognition, text processing and concept tagging techniques. We describe how this framework was successfully applied to a very large BBC radio archive. We demonstrate an application using automatically extracted tags to aid discovery of archive content.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Content Analysis and Indexing

General Terms

Algorithms, Design

Keywords

Linked Data, Concept Tagging, Speech Processing

1. INTRODUCTION

The BBC (British Broadcasting Corporation) has broadcast radio programmes since 1922 and TV since 1932. Over the years, it has accumulated a very large archive of TV and radio programmes.

Currently all programmes within the BBC archive that are likely to be reused (either through re-broadcast or as clips made available online) are manually classified. Creating this metadata by hand is an expensive process in terms of time and resources; a detailed analysis of a 30 minute programme can take a professional archivist 8 to 9 hours. Moreover, as this data is geared towards professional reuse, it is often not appropriate for driving user-facing systems — it is either too shallow (not all programmes are being classified) or too deep (information about individual shots or rushes). Also, the coverage of the catalogue is not uniform across the BBC's archive, for example it excludes the BBC World Service, which has been broadcasting since 1932.

Over the last couple of years, parallel efforts have been made to improve cross-linking between programmes and other items of interest on the BBC website. A significant part of this effort relies on tagging. A tagging tool is used by editorial teams to associate a number of Linked Data web identifiers with any resource made available on the website. In order to bootstrap this tagging process, the editor can choose from a number of tags that have been extracted from the textual content of the page that is being annotated. The entire tagging process is described in more detail in [6].

The tags are then used for a variety of use-cases across the BBC website:

- Building navigation badges, providing a consistent way of summarising topics and navigating through them across the BBC web site[1];
- Building aggregations of resources across the BBC website (topic pages)[2];
- Building aggregations of resources within a single domain[3];
- Deriving links across domains (e.g. from a Wildlife Finder species page[4] to programmes that are tagged with that species);
- Deriving links to other websites also linking to Linked Data web identifiers.

This process of manual tagging is naturally very time-consuming, and with the emphasis on delivering new content, would take considerable time to apply to the entire archive. This problem is compounded by the lack of availability of textual meta-data for a significant percentage of the archive which prevents a similar bootstrapping of the tagging process. An option for solving this problem is to use the audio content of the programmes for suggesting tags.

There has been a number of attempts at trying to automatically classify the BBC archive. The THISL system [1]

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

[1] See for example the footer of `http://www.bbc.co.uk/programmes/b014m55k`, last accessed November 2011

[2] See for example `http://www.bbc.co.uk/search/world_war_ii`, last accessed November 2011

[3] See for example `http://www.bbc.co.uk/programmes/topics/world_war_ii`, last accessed November 2011

[4] See for example `http://www.bbc.co.uk/nature/life/Lion`, last accessed November 2011

applies an automated speech recognition system (ABBOT) on BBC news broadcasts and uses a bag-of-words model on the resulting transcripts for programme retrieval. The Rich News system [5] also uses ABBOT for speech recognition. It then segments the transcripts using bag-of-words similarity between consecutive segments using Choi's C99 algorithm [3]. For each segment, a set of keyphrases is extracted and used, along with the broadcast date of the programme, to find content within the BBC News website. Information associated with retrieved news articles is then used to annotate the topical segment. Recent work at the BBC classifies the mood of archived programmes using their theme tunes [4] and ultimately intends to help users browse the archive by mood.

In this paper we describe a framework to automatically assign Linked Data web identifiers to programmes, by analysing the audio content of a programme. We describe how that framework was successfully applied to the entire World Service archive. We also describe the Tellytopic application, making use of the automatically extracted tags to aid the discovery of archive content.

Tag	Score
Programme 1	
d:Benjamin_Britten	0.09
d:Music	0.054
d:Gustav_Holst	0.024
Programme 2	
d:Revolution	0.037
d:Tehran	0.032
d:Ayatollah	0.025
Programme 3	
d:Hepatitis	0.288
d:Vaccine	0.129
d:Medical_research	0.04

Table 1: Example of automatically generated tags and associated scores. Programme 1 is a 1970 profile of the composer Gustav Holst. Programme 2 is a 1983 profile of the Ayatollah Khomeini. Programme 3 is a 1983 episode of the Medical Programme.

2. AN ALGORITHM FOR AUTOMATED TAGGING OF SPEECH AUDIO

In the following, we describe a workflow which can be used to bootstrap the tagging process of large archives of radio content.

We use the open source CMU Sphinx-3 software, with the HUB4 acoustic model [11] and a language model extracted from the Gigaword corpus[5]. The resulting transcripts are very noisy and have no punctuation or capitalisation, which means off-the-shelf concept tagging tools perform badly on them. We therefore design an alternative concept tagging algorithm.

We identify candidate terms in the transcripts. We start by generating a list of web identifiers used by BBC editors to tag programmes. Those web identifiers identify people, places, subjects and organisations within DBpedia [2]. For each of those identifiers, we dereference them and get their label from their `rdfs:label` property. We strip out any disambiguation string from the label, and apply the Porter Stemmer algorithm [10] to it in order to get to a corresponding term. We look for those terms in the automated transcripts, after applying the same stemming algorithm to them. The output of this process is a list of candidate terms found in the transcripts and a list of possible corresponding DBpedia web identifiers for them.

In order to disambiguate and rank candidate terms, we use an approach inspired by the Enhanced Topic-based Vector Space Model proposed in [7] and further described and evaluated in [9]. We consider the subject classification in DBpedia, derived from Wikipedia categories and encoded as a SKOS [8] model. We start by constructing a vector space for those SKOS categories, capturing hierarchical relationships between them. Two categories that are siblings will have a high cosine similarity. Two categories that do not share any ancestor will have a null cosine similarity. The further away a common ancestor between two categories is, the lower the cosine similarity between those two categories will be. We implemented such a vector space model within our RDFSim project[6]. We consider a vector in that space for each DBpedia web identifier, corresponding to a weighted sum of all the categories attached to it. We then construct a vector modelling the whole programme, by summing all vectors of all possible corresponding DBpedia web identifiers for all candidate terms. Web identifiers corresponding to wrong interpretations of specific terms will account for very little in the resulting vector, while web identifiers related with the main topics of the programme will overlap and add up. For each ambiguous term, we pick the corresponding DBpedia web identifer that is the closest to that programme vector. We then rank the outputted web identifiers by considering their TF-IDF score and their distance to the programme vector.

We end up with a ranked list of DBpedia web identifiers, for each programme. Some examples of the top three tags and their associated scores, obtained using an implementation of the above algorithm in Python named KiWi, are given in Table 1.

3. PROCESSING THE WORLD SERVICE ARCHIVE

In recent years the BBC World Service have begun to digitise their archive of radio programmes. The archive currently holds aroud 70,000 programmes, which amounts to about two and a half years of continuous audio.

We want to process this entire archive with the algorithm described in Section 2. However, given that the transcription process is slower than real-time on commodity hardware, we developed an infrastructure to process the entire archive in a reasonable time.

We separated each step of the workflow into independent, self-contained applications, or "workers". Each worker takes input in the form of the results of the previous step of the workflow, and produces output to be given to the next step of the workflow. We developed the following workers:

- A worker to decode and downsample programmes;

[5] See `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05`, last accessed November 2011

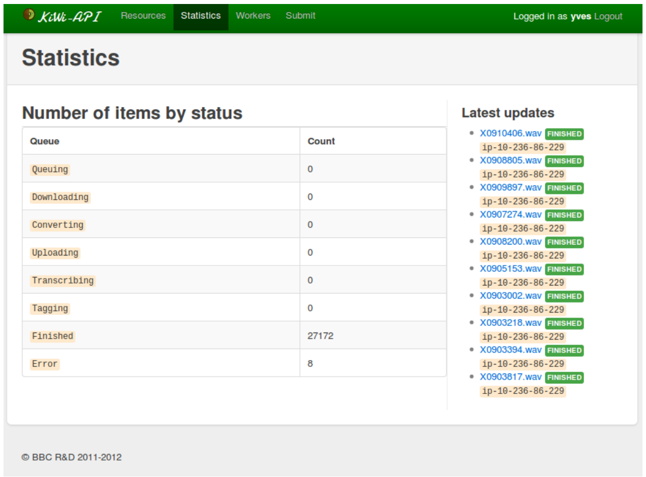
[6] See `https://github.com/bbcrd/rdfsim`, last accessed February 2012

- A worker to upload the resulting audio files to shared cloud storage;
- A worker to transcribe audio files;
- A worker to identify and disambiguate tags from programme transcripts;
- A worker to rank tags.

With such a setup, individual workers can easily be deployed to machine instances on a cloud infrastructure. We also configured a message-queueing system to allow workers to assign tasks to one-another. In order to control and monitor the system, we developed an HTTP interface which has direct access to a storage interface and to the message-queueing system. Workers update the calculated metadata and status of each processed file with a PUT request to the interface, in JSON format. Users of the system can add new files through a simple POST request, and can monitor the status or metadata of each file through a GET request. A capture of the statistics page of that interface is given in Figure 1.

With this infrastructure in place, we processed the entire World Service archive in weeks instead of years, for a predetermined cost, and generated a collection of ranked Linked Data tags for each World Service programme.

An evaluation of the resulting tags was performed by comparing them to tags manually applied by BBC editors. While the results are not perfect, they are good enough to efficiently bootstrap the tagging process.



Statistics

Number of items by status

Queue	Count
Queuing	0
Downloading	0
Converting	0
Uploading	0
Transcribing	0
Tagging	0
Finished	27172
Error	8

Latest updates

- X0910406.wav FINISHED ip-10-236-86-229
- X0908805.wav FINISHED ip-10-236-86-229
- X0909897.wav FINISHED ip-10-236-86-229
- X0907274.wav FINISHED ip-10-236-86-229
- X0908200.wav FINISHED ip-10-236-86-229
- X0905153.wav FINISHED ip-10-236-86-229
- X0903002.wav FINISHED ip-10-236-86-229
- X0903218.wav FINISHED ip-10-236-86-229
- X0903394.wav FINISHED ip-10-236-86-229
- X0903817.wav FINISHED ip-10-236-86-229

© BBC R&D 2011-2012

Figure 1: The statistics page of the KiWi API, used to manage the automated tagging of large archives

4. TELLYTOPIC

In addition to this evaluation of automatically generated tags, the BBC worked with the metadata and personalisation specialists MetaBroadcast[7] to build a prototype user interface for browsing the generated tags and their associated programmes, called Tellytopic.

Tellytopic is a web application built on top of Atlas[8], MetaBroadcast's open-source TV and radio metadata platform. For this prototype, the following datasets were loaded into Atlas:

- The `bbc.co.uk/programmes` data, including editorially added tags;
- A playable archive of BBC content broadcast since mid-2007;
- The World Service radio archive;
- The automatically extracted tags.

One of the major features of Atlas is its programme equivalence matching. This was used to match the three different datasets to produce metadata, tags and media locations for each programme based on a combination of all available data.

Tellytopic allows browsing programmes from `bbc.co.uk/programmes` (both TV and radio) and the World Service radio archive and navigating between them using a combination of tags produced automatically and assigned editorially. For example, the 'Suffolk' page in Tellytopic is captured in figure 2, and includes recent content and content from within the World Service radio archive. Users can then watch or listen to the media from one of the ingested datasets, depending on where and when it is currently available.

Tellytopic provides a useful, interesting and unique first usage of this data that demonstrates the power and value of automatic linking and tagging to a wider audience within the BBC and elsewhere. Users can navigate recent content available on the BBC website and discover related archive content. It is also used by the project team to visually explore the results of automatic tagging, helping to inform future changes to the underlying algorithms and the relationship between editorially selected and automated tags.

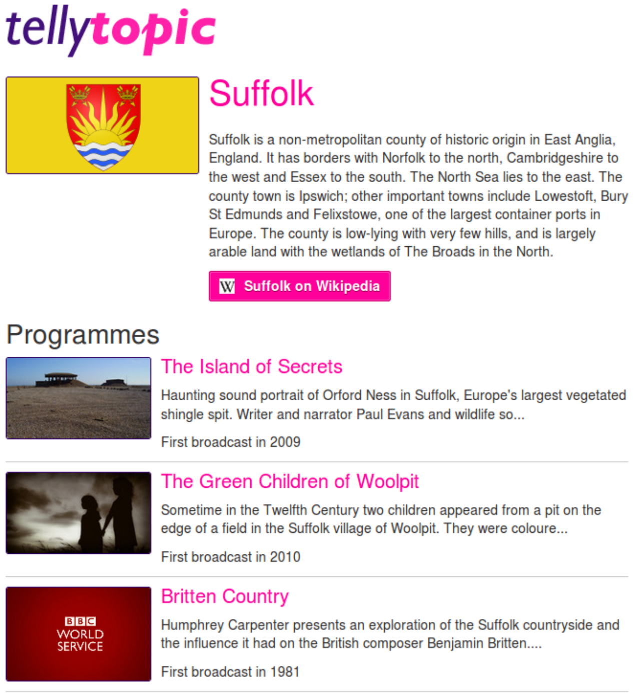
5. CONCLUSION AND FUTURE WORK

In this paper, we described an automated system for tagging speech radio programmes with web identifiers from the Linked Data cloud. We outlined an algorithm for automated tagging of speech audio, and described how this algorithm was applied to a large radio archive. We described an application of the resulting tags, TellyTopic, which allows users to navigate between an archive dataset and recent, manually annotated, programmes.

Future work includes creating an editorial interface to enable editors and the public to edit and approve the list of automatically derived tags, perhaps integrated within the Tellytopic application. In particular, we are currently investigating the user experience issues around publishing archive with data we know is not accurate, whether extracted from legacy systems or automatically generated. We also want to try and incorporate more data (synopsis, broadcast dates, etc.) in the automated tagging process when this data is available. Another potential area of work is to improve the results of the automated speech recognition step, by creating a specific acoustic model for British English, and a language model built from programme transcripts.

[7] `http://metabroadcast.com`, last accessed November 2011

[8] `http://atlas.metabroadcast.com/`, last accessed November 2011



Suffolk

Suffolk is a non-metropolitan county of historic origin in East Anglia, England. It has borders with Norfolk to the north, Cambridgeshire to the west and Essex to the south. The North Sea lies to the east. The county town is Ipswich; other important towns include Lowestoft, Bury St Edmunds and Felixstowe, one of the largest container ports in Europe. The county is low-lying with very few hills, and is largely arable land with the wetlands of The Broads in the North.

Programmes

The Island of Secrets

Haunting sound portrait of Orford Ness in Suffolk, Europe's largest vegetated shingle spit. Writer and narrator Paul Evans and wildlife so...

First broadcast in 2009

The Green Children of Woolpit

Sometime in the Twelfth Century two children appeared from a pit on the edge of a field in the Suffolk village of Woolpit. They were coloure...

First broadcast in 2010

Britten Country

Humphrey Carpenter presents an exploration of the Suffolk countryside and the influence it had on the British composer Benjamin Britten....

First broadcast in 1981

Figure 2: A Tellytopic aggregation page of programmes tagged with the Suffolk web identifier, including both recent content and content from the World Service radio archive

6. ACKNOWLEDGEMENTS

The research for this paper was conducted as part of the Automatic Broadcast Content Interlinking Project (ABC-IP). ABC-IP is a collaborative research and development initiative between the British Broadcasting Corporation and MetaBroadcast Ltd, supported with grant funding from the UK Technology Strategy Board under its 'Metadata: increasing the value of digital content (mainstream projects)' competition from September 2010.

7. REFERENCES

[1] Dave Abberley, David Kirby, Steve Renals, and Tony Robinson. The THISL broadcast news retrieval system. In *Proc. ESCA Workshop on Accessing Information In Spoken Audio*, 1999.

[2] Soren Auer, Christian Bizer, Jens Lehmann, Georgi Kobilarov, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the International Semantic Web Conference*, Busan, Korea, November 11-15 2007.

[3] Freddy Y. Y. Choi. Advances in domain independent linear text segmentation. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 2000.

[4] Sam Davies, Penelope Allen, Mark Mann, and Trevor Cox. Musical moods: A mass participation experiment for affective classification of music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 2011.

[5] Mike Dowman, Valentin Tablan, Hamish Cunningham, and Borislav Popov. Web-assisted annotation, semantic indexing and search of television and radio news. In *WWW '05 Proceedings of the 14th international conference on World Wide Web*, 2005.

[6] Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Chris Bizer, and Robert Lee. Media meets semantic web - how the BBC uses DBpedia and linked data to make connections. In *Proceedings of the European Semantic Web Conference In-Use track*, 2009.

[7] D. Kuropka. *Modelle zur Repräsentation natürlichsprachlicher Dokumente - Information-Filtering und -Retrieval mit relationalen Datenbanken*. Logos Verlag, 2004. ISBN: 3-8325-0514-8.

[8] Alistair Miles, B. Matthews, M. Wilson, and D. Brickley. SKOS core: Simple knowledge organisation for the web. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (DC-2005),*, pages 5–13, Madrid, 2005.

[9] Artem Polyvyanyy. Evaluation of a novel information retrieval model: eTVSM. Master's thesis, Hasso Plattner Institut, 2007.

[10] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[11] Kristie Seymore, Stanley Chen, Sam-Joo Doh, Maxine Eskenaziand Evandro Gouvea, Bhiksha Raj, Mosur Ravishankar, Ronald Rosenfeld, Matthew Siegler, Richard Sternane, and Eric Thayer. The 1997 CMU sphinx-3 english broadcast news transcription system. In *Proceedings of the DARPA Speech Recognition Workshop*, 1998.