Automatic Metadata Generation Through Analysis Of Narration Within Instructional Videos

Joseph Rafferty[1], Chris Nugent[1], Jun Liu[1] and Liming Chen[2]

[1] *School of Computing and Mathematics, University of Ulster*

[2] *School Computer Science and Informatics, De Montfort University*

Email: rafferty-j@email.ulster.ac.uk, {cd.nugent, j.liu} @ulster.ac.uk, liming.chen@dmu.ac.uk,

Abstract.
Current activity recognition based assistive living solutions have adopted relatively rigid models of inhabitant activities. These solutions have some deficiencies associated with the use of these models. To address this, a goal-oriented solution has been proposed. In a goal-oriented solution, goal models offer a method of flexibly modelling inhabitant activity. The flexibility of these goal models can dynamically produce a large number of varying action plans that may be used to guide inhabitants. In order to provide illustrative, video-based, instruction for these numerous actions plans, a number of video clips would need to be associated with each variation. To address this, rich metadata may be used to automatically match appropriate video clips from a video repository to each specific, dynamically generated, activity plan. This study introduces a mechanism of automatically generating suitable rich metadata representing actions depicted within video clips to facilitate such video matching. This performance of this mechanism was evaluated using eighteen video files; during this evaluation metadata was automatically generated with a high level of accuracy.

Keywords: Assistive Living, Automated Speech Recognition, Metadata, Ontology, Parsing, Smart Environments, Video.

Introduction

The global population is aging and as a result is developing an uneven demographic composition. It is projected that by 2050 over 20% of the population will be aged 65 or over [1]. This growth in aging population is expected to produce an increase in age related illness and so will place additional burdens on healthcare infrastructure [2].

The use of technological solutions to support independent living offers promise to alleviate a subset of these aging related problems [3]. This may come in the form of a Smart Home (SH). SHs are residential environments which have been augmented with a variety of technologies in order to provide assistive function.

These technologies include; a sensor deployment in order to monitor inhabitants, a processing layer to support some form of activity recognition or prediction and actuators to provide assistance [3, 4]. Current SHs typically provide assistance in the form of prompting systems, monitoring of behavioural trends and remote assessment of vital signs. In SHs, prompting systems offer guidance for inhabitants as determined by the processing layer. Prompts may consist of video, audio or text, or a combination thereof. Of these types of prompt, video incorporating guiding narration provides a promising method of providing informative and relatable instruction [3, 5].

This study introduces a novel mechanism for analysing video clips in order to generate metadata. This metadata is to be used in support of the provision of dynamic instruction within a goal driven SH paradigm. This remainder of the paper is arranged as follows: an overview of related work is presented; the approach used in this study is detailed; an evaluation of the approach is provided and a conclusion is offered.

Related work

Currently, a number of SHs exist which provide assistance for inhabitants by offering guidance by prompting [3, 4, 6–8]. The prompting systems in these environments include use of audio, text or video instruction. Of these type of prompts, video based instruction provides detailed and relatable instruction for inhabitants of a SH.

Current SHs that offer video based guidance do so in a relatively static manner. This is due to the use of inflexible structures used to model activities [9]. To address the inflexibility of modelling activities, SHs focusing on inhabitant goals in place of traditional activity models have been proposed [9]. Goals contain action plans, which detail how to achieve a specific task such as placing a teabag in a cup, or boiling water. These goals, and their action plans can then be inherited by other goals through the use of a sub-goal property. This inheritance allows a multitude of goals to be modelled from a combination of sub-goals and so supports variations of activities and flexibility with their performance. An example of this inheritance is illustrated in Fig. 1, where 7 goals containing action plans, as represented by atomic actions, are combined in various ways to produce 16 distinct goals.

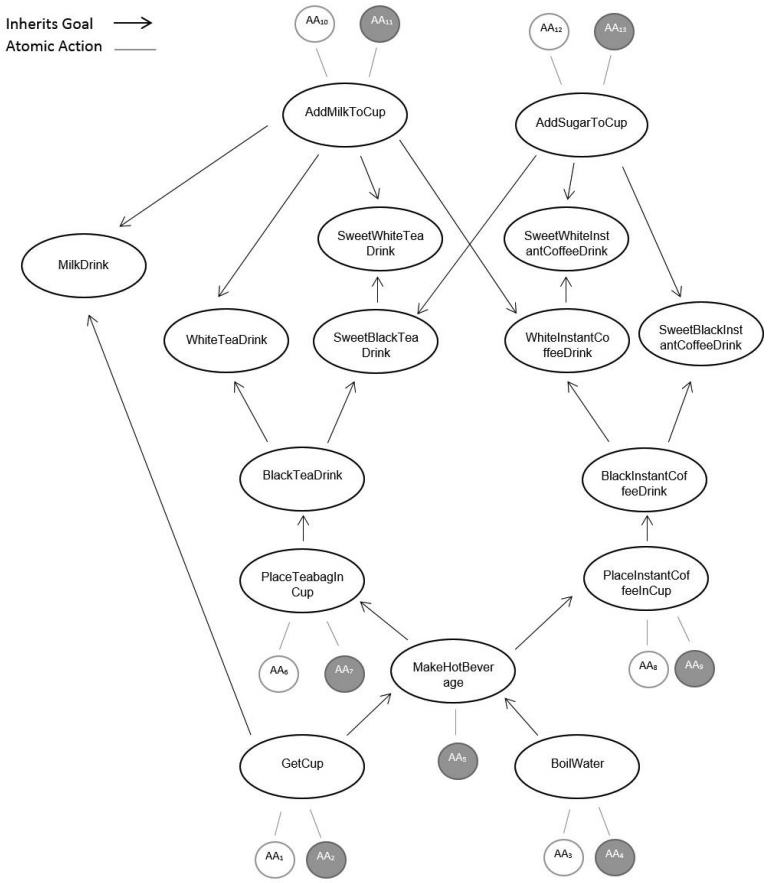

Fig. 1. The flexibility of modelling activities using inhabitant goals, in this illustration 7 goals that have associated action plans, as represented by atomic actions, may be combined to produce 16 distinct goals.

To date, video instruction for activity guidance in SHs is provided in a static manner. Such static assignment is not compatible with a goal-oriented approach due to the added flexibility offered by this approach. Video based guidance within a goal driven approach, therefore, requires a flexible mechanism to determine the most suitable candidate video from a repository. In order to identify a suitable match from a repository, accurate and complete metadata is required. Such metadata can subsequently be used with a selection mechanism to nominate a suitable video clip.

Currently, metadata for such video clips are usually provided manually, which may lead to incomplete or incorrect records. Additionally, production of complete metadata which provides indication of all tasks that are depicted in a video would require a large amount of effort [10]. Automatic generation of such metadata will provide a method to gain consistent, accurate and complete information about the tasks within the depicted in a video.

Currently, automatic video based metadata generation mainly involves analysis of object interactions within a scene, performing analysis on textual elements within videos or analysis of content using statistical analysis [11–14]. Audio based metadata generation mostly focuses on sound analysis, with limited work incorporating Automated Speech Recognition systems (ASR) which convert speech to text [14–16]. These methods do not provide a suitable method of producing activity annotations, as they cannot currently identify a set of goal actions in such video clips.

In addition, the majority of these approaches require training with a dataset beforehand, resulting in a cold start problem [11–14]. Additionally, computer vision techniques require a significant amount of computation over ASR based analysis. As the videos to be presented by such a goal driven SH will have clear, concise narration it would be more appropriate to focus on analysis the audio aspect of such video instruction.

Automatic metadata generation through analysis of narration within instructional videos

In this study, an annotation method capable of generating rich metadata for video clips has been created and implemented within an evaluation platform called ABSEIL (Audio BaSEd Instruction profiLer). This platform is intended to work in conjunction with the video repository generated by the Personal IADL Assistant (PIA) project[1]. The goal of the PIA project is to assist older individuals by offering guidance with Instrumental Activities of Daily Living (IADL) [17]. IADLS are those activities that are required to live independently within a community, for example meal preparation.

In PIA, caregivers record videos which contain an accompanying, detailed, narration explaining how to achieve a task associated with an IADL. These videos are then associated with NFC tags[2] which are affixed to items in an environment. Those in need of assistance can use a compatible smart phone or tablet to interact with such an NFC tag in order to play the associated video. To offer effective support, the videos within the PIA repository should therefore be of assured quality and contain clear narration. As such, PIA will provide a source of high quality instruction suitable for guiding inhabitants of a smart home. Further information about the PIA project is available in [18].

In the devised approach, videos are converted to audio clips. The audio clips are then sent to a black-box ASR [19] which returns a transcription. This transcription is processed to identify goal actions from a goal repository to determine if any are present within the narration. Due to personalisation, terms used in narration and modelling of the goal actions may differ in specific words used but still refer to semantically compatible variations. As such, compatible, alternative, variations of these actions are created using a semantic lexicon and homophone dictionary. A depiction of this method is presented in Fig. 2.

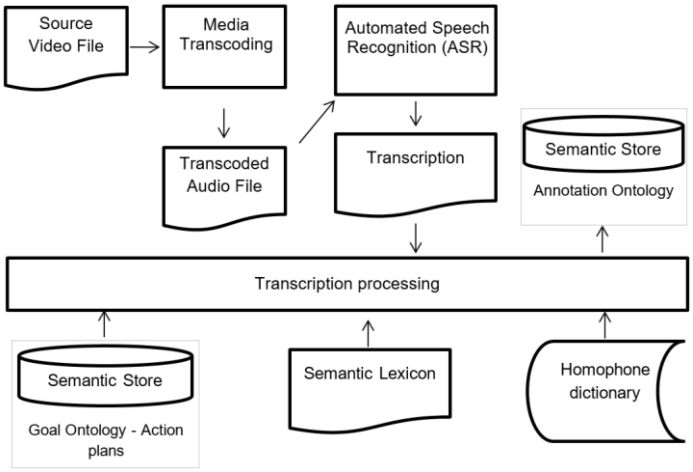

Fig. 2. An overview of the automatic metadata generation process.

[1] Personal IADL Assistant, PIA – EU AAL Funded Research Project (AAL-2012-5-033), available at: http://www.pia-project.org/

[2] Near Field Communication – A short range contactless communication technology

ABSEIL uses the FFMPEG suite [20] to transcode a candidate video file into a set of audio files. These audio files vary in channel count (mono or stereo), compression (AAC, FLAC, MP3, and PCM), sample size (8bit, 16-bit, 32-bit), sample rates (e.g. 8000Hz, 22000Hz 44100Hz) and container format (.MP4, .FLAC, .MP3 and .WAV). Creation of this set of files allows audio clips to be submitted to a range of ASR systems.

When the audio files have been produced, the FLAC version is then streamed to the Google Speech API (GSAPI) [21], an ASR system. GSAPI was chosen over alternative, available, ASR solutions (Apple, Dragon, Windows Speech) due to its better performance during evaluation of systems for this application. This evaluation was achieved by selecting 3 videos, from the PIA project, manually creating a transcription of this narration and comparing the accuracy of the manual transcription to those produced by ASR systems. The Sørensen–Dice coefficient was used as a string similarity metric [22], the equation for this metric is shown in (1). This coefficient is best suited to comparing large, generic, variable, string sets such as these without a bias (unlike other metrics such as Jaccard), as such it was chosen for this purpose [23, 24].

$$S = \frac{2C}{A+B} = \frac{2|A \cap B|}{|A|+|B|} \quad (1)$$

The result of this comparison is presented in Table 1. It should be noted that ASR systems may be trained, tailoring the ASR mechanism towards recognising a specific voice. Given that video clips in the PIA repository would have multiple narrators, this training process would prove to be detrimental to the generic ASR performance required, for example if trained to a female voice ASR performance would have decreased performance with male voices. Subsequently, ASR systems were not trained in this evaluation (where possible). In this assessment the large performance gap between GSAPI and others can be explained by the different operational models of these systems. GSAPI is a hosted service that constantly learns from google data mining , media processing and search operations. The others are static in place systems that have traditional releases that are relatively static after their inception and distubition.

Table 1. An assment ASR systems, coparing automated transcriptions to a manual transcription.

ASR system	**Accuracy of transcriptions**			
	Video 1	*Video 2*	*Video 3*	*Average*
*Apple**	0% - No Result			0%
Dragon†	0% - No Result			0%
*Google (GSAPI)**	70.0%	57.14%	68.57%	65.24%
Windows Speech †	26.84%	22.38%	33.52%	27.59%

Online Systems are indicated by *, untrained systems are indicated by †

In this approach, a repository of modelled goals for a SH exists. This repository may contain inhabitant goals such as *MakeCoffee,* which may also reference sub-goals, such as *GetCup*. Goals have an associated action plan which contains a number of atomic actions, these are steps required to achieve the goal. The atomic actions from all goals within the ontology are used to generate a four search sets. These sets are used to process the transcription, identifying matches. These sets are *direct, homophone substitution, synonym substitution* and *homophone/synonym substitution*.

The evaluated ASR systems are black-box mechanisms and so offer limited insight into, or influence over, the transcription process. This introduces issues when homophones are

encountered. Homophones are words which may be phonetically confused, for example '*birth'* and '*berth*'.

All of the evaluated ASR systems had issues with selecting the correct homophones for a transcription. As such it was necessary to introduce some mechanism to account for these errors. Contemporary works within ASR correct for these homophones within the statistical/machine learning core of the ASR process [19, 25]. Due to the black box use of these ASR systems, an alternative method of substitution needs be implemented. Instead, this correction needs to be applied to the terms used to search the produced transcriptions. In this approach, substitution produces combinations of words from a pre-existing homophone dictionary [26]. The homophone dictionary used was a compilation of the Australian and American homophone lists produced by the Summer Institute of Linguistics. These combinations are placed in a *homophone substitution* search set. For example the set generated from the atomic action *PourWater* is [*pore water, pour water, poor water*]. This set provides some opportunity for correction of errors related to homophones.

The atomic actions provided by the goal plan provide a description of the task represented by that action, for example *PlaceCoffeeInCup*. The exact words used to depict these actions are specified by the person modelling that goal. This introduces an issue when terms used in narration of illustrative videos use alternative, semantically compatible, words to those used in the atomic actions.

For example, an utterance of "Place coffee in mug" in narration of an instructive video could be used to describe an action which is effectively compatible with the *PlaceCoffeeInCup* atomic action. In place of the word '*cup'* some of its synonyms could be used, some of these are [*cup, mug, teacup*]. To cater for this, a lexicon that can identify and generate a listing of synonyms was used to produce *synonym substitution* search sets. In the ABSEIL platform, the chosen semantic lexicon was WordNet 3.1 [27]. WordNet is maintained by Princeton University and is one of the foremost lexical databases for the English language. In WordNet, words are stored within synsets that act as the source of semantically compatible words in ABSEIL.

In order to correct for instances where both homophone errors and use of semantically compatible words occur it is necessary to create a search set that contains these combinations, the *homophone/synonym substitution* search set. An example of such a combination from the action *PourWaterInCup* could include [*poor water in mug*, *poor water in mug*]. The combinations in this set provide useful alternatives to atomic actions that are present in such narrated videos.

Once the search sets have been generated they are used to search the body of the transcription to identify utterance of their terms. During the search process, words within search terms are given a four word window between them. This allows atomic actions such as *OpenCupboard* to be found when alternative variations are uttered, for example "Open the top cupboard". In the evaluation implementation this search is performed by the Apache Lucene text processing engine [28]. This text processing engine was chosen opposed to a dedicated Natural Language Processing (NLP) toolkit as it offers better performance when processing large bodies of text, such as transcriptions, especially when using a large number of search terms as is the case in this approach.

When matches from the four search sets are discovered they are stored in the video action annotation ontology for the analysed video. In addition to any matches, this video ontology stores some metadata about videos; a unique identifier, an optional title and an optional description. The unique identifier is a SHA512 hash of the original video file facilitating a

means to associate the metadata to the file. By default the optional video title and descriptions are retrieved from the PIA repository, where possible.

The ontology contains four classes to hold matched terms from the four search sets. These are the *DirectTerms, HomophoneTerms, SynonymTerms* and *HomophoneSynonymTerms* classes. All these classes contain the same set of data properties: *DepictedAction, TimeStamp and Duration*. In the current ABSEIL platform, the *TimeStamp* and *Duration* properties are not used, their inclusion is to provide support for future iterations where the timestamp of each action is recorded. The video action annotation ontology is depicted in Fig. 3 and Fig. 4.

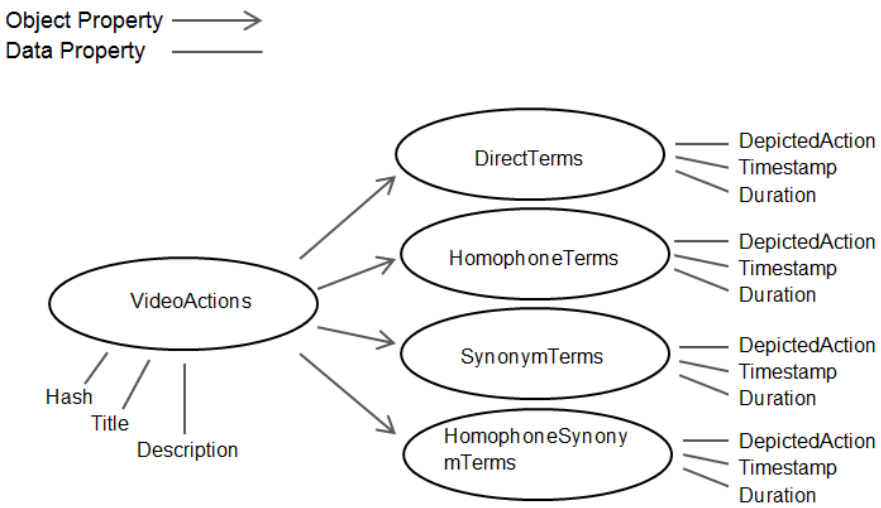

Fig. 3. The classes and object/data properties of the video action annotation ontology. As shown as a hierarchy of concepts (a).

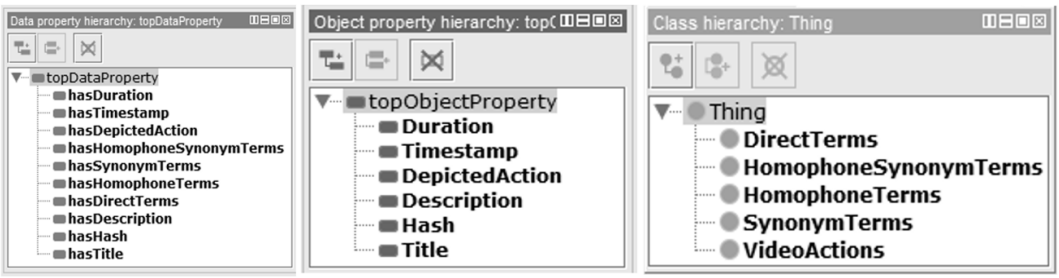

Fig. 4. The classes and object/data properties of the video action annotation ontology. As shown the Protégé ontology engineering tool (b).

Evaluation

To evaluate the performance of this automatic annotation generation method it was implemented in the form of the ABSEIL system. During evaluation, eighteen instructional videos were evaluated. Eight of which were from the PIA project and ten were narrated, PIA-Style, instructional videos. During evaluation, manually generated metadata was produced for these videos using the atomic action set that was extracted from a “Making Instant Coffee” goal and its subgoals. The videos were analysed with the ABSEIL platform and its accuracy was compared to the manually created metadata.

The PIA video set contained a single video that involved the steps of making coffee and so this set was used to evaluate incorrect profiling of video clips, where metadata would be generated without any relevant content. The independently narrated videos covered a range of beverage making tasks and were used to determine the success of the approach. The results of testing the implementation are presented in Table 2. In this evaluation each of the four search sets were assessed and averaged, erroneously identified actions are noted as false positives.

Table 2. The results of evaluating the accuracy of metadata produced by the automatic metadata method compared to manually generated metadata. False positives are indicated in brackets.

Video Source	**Accuracy of generated annotation**			
	DirectTerms	*HomophoneTerms*	*SynonymTerms*	*HomophoneSynonymTerms*
PIA (8 videos)	87.5 % (0)	87.5 % (0)	87.5 % (0)	87.5 % (0)
Independently Narrated (10 videos)	66.86% (0)	73.97 % (0)	79.93% (2.9)	82.59 % (2.7)

As shown in this evaluation, the devised method shows promise for use in automatically generating the metadata required guidance provision for a goal driven SH.

In some instances semantically compatible words were used in place of those specified in the action plan and were not discovered by the processing. In order to remedy this, additional sources of synonyms may be introduced.

False positives present were nonsensical phrases involving additional prepositions or combinations. Such nonsensical phrases would not be present in candidate videos intended to provide clear instruction.

Additional issues were encountered when a narrator referred to a previous object as “it”. Such utterances can handled by the incorporation of a more advanced NLP toolkit.

Conclusion

In this paper a method of automatically generating metadata for video files was presented. This metadata represents actions depicted within the narration of an instructional video file. This metadata is generated with the objective of providing the basis of offering assistance within a goal-driven SH. This method has been implemented in an evaluation platform, named ABSEIL, and shows promise in automatically generating metadata. Future work will include producing assistance provisioning mechanisms which will leverage this rich metadata to automatically provide illustrative guidance that is best suited to a given inhabitant goal.

Acknowledgments.
This work has been conducted in the context of the EU AAL PIA project (AAL-2012-5-033).
The authors gratefully acknowledge the contributions from all members of the PIA consortium.

References

1. De Luca, d’Alessandro E., Bonacci, S., Giraldi, G.: Aging populations: the health and quality of life of the elderly. Clin. Ter. 162, e13 (2011).
2. United Nations: World Population Ageing 2009 (Population Studies Series). (2010).

3. Acampora, G., Cook, D.J., Rashidi, P., Vasilakos, A. V: A Survey on Ambient Intelligence in Health Care. Proc. IEEE. Inst. Electr. Electron. Eng. 101, 2470–2494 (2013).
4. Chen, L., Hoey, J., Nugent, C.D., Cook, D.J., Yu, Z.: Sensor-Based Activity Recognition. IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev. 1–19 (2012).
5. Lapointe, J., Bouchard, B., Bouchard, J.: Smart homes for people with Alzheimer's disease: adapting prompting strategies to the patient's cognitive profile. Proc. 5th Int. Conf. PErvasive Technol. Relat. to Assist. Environ. 3, (2012).
6. Chan, M., Estève, D., Escriba, C., Campo, E.: A review of smart homes- present state and future challenges. Comput. Methods Programs Biomed. 91, 55–81 (2008).
7. Cook, D.J., Das, S.K.: How smart are our environments? An updated look at the state of the art. Pervasive Mob. Comput. 3, 53–73 (2007).
8. Mihailidis, A., Boger, J.N., Craig, T., Hoey, J.: The COACH prompting system to assist older adults with dementia through handwashing: an efficacy study. BMC Geriatr. 8, 28 (2008).
9. Rafferty, J., Chen, L., Nugent, C.: Ontological Goal Modelling for Proactive Assistive Living in Smart Environments. Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction. pp. 262–269 (2013).
10. Filippova, K., Hall, K.: Improved video categorization from text metadata and user comments. SIGIR '11 Proc. 34th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. 835–842 (2011).
11. Papadopoulos, D.P., Kalogeiton, V.S., Chatzichristofis, S. a., Papamarkos, N.: Automatic summarization and annotation of videos with lack of metadata information. Expert Syst. Appl. 40, 5765–5778 (2013).
12. Ballan, L., Bertini, M., Bimbo, A., Seidenari, L., Serra, G.: Event detection and recognition for semantic annotation of video. Multimed. Tools Appl. 51, 279–302 (2010).
13. McCloskey, S., Davalos, P.: Activity detection in the wild using video metadata. Pattern Recognit. 3140–3143 (2012).
14. Perea-Ortega, J.M., Montejo-Ráez, A., Martín-Valdivia, M.T., Ureña-López, L.A.: Semantic tagging of video ASR transcripts using the web as a source of knowledge. Comput. Stand. Interfaces. 35, 519–528 (2013).
15. Metze, F., Ding, D., Younessian, E., Hauptmann, A.: Beyond audio and video retrieval: topic-oriented multimedia summarization. Int. J. Multimed. Inf. Retr. 2, 131–144 (2013).
16. Papadopoulos, D.P., Kalogeiton, V.S., Chatzichristofis, S. a., Papamarkos, N.: Automatic summarization and annotation of videos with lack of metadata information. Expert Syst. Appl. 40, 5765–5778 (2013).
17. Lawton, M., Brody, E.: Instrumental Activities of Daily Living Scale (IADL). (1988).
18. Rafferty, J., Nugent, C., Chen, L., Qi, J., Dutton, R., Zirk, A., Boye, L.T., Kohn, M., Hellman, R.: NFC based provisioning of instructional videos to assist with instrumental activities of daily living. Engineering in Medicine and Biology Society (2014).
19. Mehla, R., Aggarwal, R.: Automatic Speech Recognition: A Survey. Int. J. Adv. Res. Comput. Sci. Electron. Eng. 3, 45–53 (2014).
20. FFMPEG, https://www.ffmpeg.org/.
21. Google: Google Speech API, http://www.google.com/speech-api/v1/recognize.
22. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology. 26, 297–302 (1945).
23. Lee, L.: On the Effectiveness of the Skew Divergence for Statistical Language Analysis. AISTATS (Artificial Intell. Stat. 65–72 (2001).
24. Cohen, W.W., Ravikumar, P.D., Fienberg, S.E.: A Comparison of String Distance Metrics for Name-Matching Tasks. Proc. IJCAI-03 Work. Inf. Integr. Web. 73–78 (2003).
25. Chen, W., Ananthakrishnan, S.: ASR error detection in a conversational spoken language translation system. Acoust. Speech Signal Process. (ICASSP), 2013 IEEE Int. Conf. 7418 – 7422 (2013).
26. SIL: American English Homophones, http://www-01.sil.org/linguistics/wordlists/english/.
27. Princeton University: About WordNet., http://wordnet.princeton.edu.
28. Apache: Lucene, http://lucene.apache.org.