Automatic Detection of Nominal Entities in Speech for Enriched Content Search

Ricardo A. Calix [±], Leili Javadpour [⊦], Mehdi Khazaeli [⊦], and Gerald M. Knapp [⊦]

[±] Computer Information Technology and Graphics, 279 Gyte, Purdue University Calumet, Hammond, IN, 46323, USA

[⊦] Mechanical and Industrial Engineering, 2508 Patrick F. Taylor Hall, Louisiana State University, Baton Rouge, LA, 70803, USA
ricardo.calix@purduecal.edu; {sjavad1, marabk2, gknapp}@lsu.edu

Abstract

In this work, a methodology is developed to detect sentient actors in spoken stories. Meta-tags are then saved to XML files associated with the audio files. A recursive approach is used to find actor candidates and features which are then classified using machine learning approaches. Results of the study indicate that the methodology performed well on a narrative based corpus of children's stories. Using Support Vector Machines for classification, an F-measure accuracy score of 86% was achieved for both named and unnamed entities. Additionally, feature analysis indicated that speech features were very useful when detecting unnamed actors.

Introduction

Retrieving content from audio collections is an important topic in information retrieval. There are many audio collections on the web that may contain what a user needs but that do not contain the metadata needed to find it. As a result, new tools will be needed to bridge the gap between semantic meaning in speech content and low level speech features. When looking for audio or video files, users usually think of ideas, emotions, or people they want to see or hear about. Therefore, developing efficient technologies that can detect these semantic concepts, retrieve them, and make them available in an enriched way could improve communication and overall user satisfaction. Additionally, finding sentient entities in speech stories for content enrichment could be very useful for visualization purposes. Text-to-scene approaches such as the ones proposed in Coyne and Sproat (2001) could use this type of enrichment technique to provide an automatically generated 3-D scene of the story with the detected actors.

The goal of this work is to develop and test a methodology for sentient entity or actor detection in speech files for use in content enrichment and better retrieval. The proposed methodology enriches audio files by adding semantic meta-tags based on automatic content detection. In particular, the methodology detects sentient actors in speech stories and then enriches the content by creating XML files with inferred characteristics about the actor. The proposed methodology is performed in 4 phases. First, the speech files are transcribed using an automatic speech recognition system. In this study, this step is simulated by using an already annotated corpus. Second, the actor candidates and text features are extracted from the text transcripts. Third, the speech features from the sentence where the actor candidate is present are extracted. Finally, the feature vectors are used to train and test the model. Classification results using various supervised learning techniques as well as an analysis of the contribution of speech and text features for actor detection are presented.

Literature Review

Speech Search and Enrichment

Studies that have addressed the issue of automatic content search in speech can be classified based on the features they use into studies that use acoustic features only, studies that use Natural Language Processing (NLP) techniques in text only, and studies that combine both approaches. For instance, Bigot et al. (2010) is a study that mainly uses acoustic approaches. The study proposes a methodology for speaker role recognition in conversational speech. Their methodology tries to classify speakers into categories such as anchor, journalist or other. To achieve this, the audio files are segmented using speaker diarization algorithms to find the different speaker patterns. Once the files are segmented, temporal, acoustic, and prosodic features are extracted to perform classification. Using hierarchical supervised classification, the system achieves accuracies of 92% for speaker role detection. In contrast, Stuker et al. (2010) propose a methodology that uses both text and speech features to retrieve German news clips based on natural spoken queries. The methodology proposed in their work uses GMMs and HMMs to model and classify the segments in the audio recording. Once a new language query is transcribed using an ASR, the system uses the Okapi distance measure to find the document closest to the query.

Actor Detection

In this work, sentient actor detection refers to techniques used to identify entities that perform actions and are subject to social characteristics such as emotions. Most studies in actor detection relate to Named Entity Recognition (NER) in text as the much broader problem of nominal actor detection (for unnamed entities) is highly complex and is still an open research problem (Pang and Fan 2009a). Actors can be grouped into named entity and unnamed entity categories. Named entities have proper names such as Lucinda, Jane, and Neo. Unnamed entities do not have a proper name but are instead referenced typically in a noun phrase such as shadow knight, the wizard, the big bad wolf. Named entity recognition (NER) approaches have been addressed in the literature and have relatively mature implementations available (Chifu et al. 2008, Nguyen and Cao 2008). The work done by Johnson et al. (2011) focuses on extracting expanded entity phrases from free text using ontology structures for handling acronyms, synonyms and abbreviations. Xi and Li (2011) also presented an approach for recognizing named entities by identifying relations. Features are extracted and classification is done to find the candidate named entities. Then they use maximum entropy to classify the relations of the candidate named entities. On the other hand, unnamed entity recognition methods are less common in the literature. The main approaches for unnamed entity detection use named entity recognition followed by disambiguation. McShane (2009) and Cassimatis (2009), for instance, argue that the next generation of intelligent systems will need knowledge about the world in order to effectively detect objects and events.

Methodology

The objective of the methodology is to detect sentient entities or actors (also referred to as Nominal Entities) in speech files for content enrichment purposes. The methodology is divided into 4 phases. First, the speech files are transcribed using an ASR system. Second, the actor candidates and text features are extracted from the text transcripts. Third, the methodology extracts the speech features from the sentence speech segment where the actor candidate was found. Finally, supervised learning algorithms and the extracted actor candidates are used to train and test a classification model. Two classes are used for the classification: actors and non-actors.

Phase 1: Speech Signal Processing

Approaches in speech content search require the use of text and acoustic features. Acoustic features can be obtained directly from the signal after segmentation. Text features can be obtained by automatically transcribing the audio signal into text using an Automatic Speech Recognition (ASR) system. In this work, this step will be simulated by using the affect corpus 2.0 which has been annotated for this task (Calix and Knapp 2011; Alm 2008). This corpus consists of 89 Fairy Tales by three authors which are the Brothers Grimm, H. C. Andersen, and B. Potter. Each of the 89 stories in the corpus includes a text and audio version. These files were manually annotated and are used to extract speech features at the sentence level using Praat scripts (Boersma et al. 2005). The annotated files are stored as text files and are used to extract the semantic content. The corpus also includes annotations of the actors in each story and their presence in a given sentence. The corpus is available from LSU-NLP (2011).

Phase 2: Actor Candidate and Text Feature Extraction

In this work, actors are defined as sentient beings such as humans or fairy tale characters capable of human like behavior, in particular, speech, thought, and emotions. The methodology proposed in this phase for extracting actor candidates is divided into 2 main steps. The first step involves parsing of the text documents using state of the art taggers, parsers, and other architectures. The second step detects noun phrase candidates and extracts features for each sample. The initial pre-processing step to produce the syntactic parses uses the Stanford parser (Klein and Manning 2002). The syntactic parses produced by the Stanford parser are used for noun phrase detection (chunking). The following short sentence is used to illustrate the methodology for actor detection: "In a certain mill lived an old miller who had neither wife nor child, and three apprentices served under him."

Using the previous sentence, an example of a syntactic parse can be seen in Figure 1 where a noun phrase is: [(NP (DT an) (JJ old) (NN miller))]. This Noun Phrase is considered an actor candidate in this work because it is a sub-tree of height 3 where the height is the distance between the head NP and the leaves of the tree (i.e. the words). Special tags like NNP can specifically help to identify named entities. After pre-processing, feature extraction occurs in the second step by iterating through each story and extracting NP candidates and their respective features. The features used in this work are divided into 4 categories which are: syntactic, knowledge-based, relation to pronouns, and general context based. Each story is loaded to memory and the algorithms iterate through it sentence by sentence in sequential order. The procedure is summarized in Algorithm 1 (Figure 2) where ST is the story, "s" is the current sentence, BranchStack is a stack used to capture the POS tags that make up a branch from the root to an NP candidate, ExtractEntities is a recursive DFS traverse function for each syntactic parse, T = (V, E) is the sentence syntactic parse which is a rooted acyclic connected graph, E = {(x, y) ⎢x, y ϵ V} is the set of edges, and V are the vertices. A recursive DFS (Depth First Search) algorithm (Figure 3) was implemented to extract NP candidates and relevant features from each sentence. After a sentence is selected, the sentence parse T is extracted and is used by the recursive DFS function ExtractEntities. This function traverses each node V in the tree. This method selects as candidates all NPs of height 3 where the height is the distance from the head NP to the leaves of the tree. For each actor candidate, a branch traversal path (branch stack) is recorded.

```
Algorithm 1.   Iterate Through Each Story to Extract actor
candidates and text features
Input: Story
Output: NP candidates with text-based features
∀s ∈ ST    do
    T ← s.CurrentSentenceParse
    BranchStack ← {}
    ExtractEntities (T, BranchStack)
```

Figure 2 Algorithm 1 – Story traversal

BranchStack is a stack used to sequentially collect the tags that make up the path (or branch) from the root node to the leaves of the actor candidate. This stack allows for an actor candidate (NP tree) to be identified as being a child of a PP node, NP node, VP node, and SINV node. Knowing the branch for each NP as a feature helps to identify the subject-object relation of the NP candidate. SINV, for instance, is used for inverted declarative sentences (e.g. when the subject follows a tensed verb).

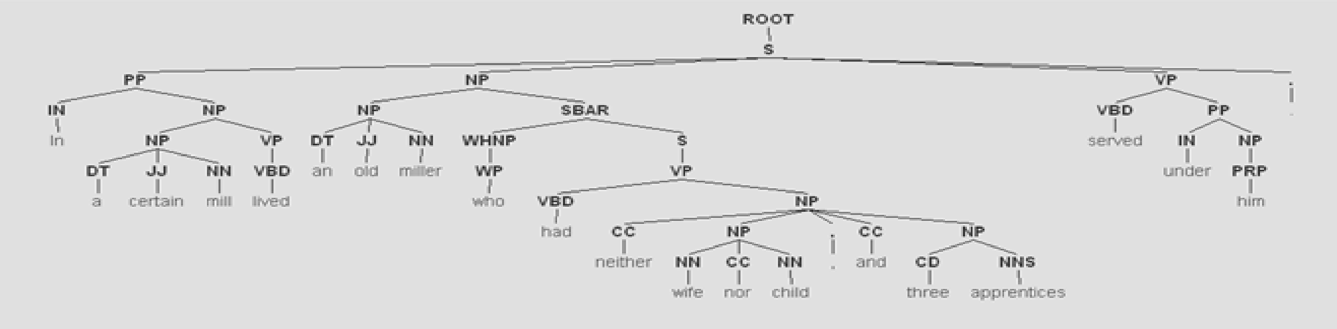

Figure 1 Syntactic parse for a input sentence from "The poor miller's boy and the cat" by the brothers Grimm

A breadth first heuristic rule was also used to detect special words that indicate if an NP is a sentient being. For example, for the chunk "said the king", the algorithm looks in the syntactic parse for sequences of tags and words (e.g. double quote followed by said the actor). The recursive DFS ExtractEntities algorithm (Figure 3) uses the function "MakeActorCandidateFeatures" to perform individual analysis for each NP candidate tree. Here syntactic, knowledge based, and other general features are extracted. Knowledge features are extracted using a call to ConceptNet (Havasi et al. 2007), which includes many concepts about different everyday types of people, things, thoughts, etc. For each entity, ConceptNet can produce many types of relations to concepts such as HasA, IsA, CapableOf, Desires, HasProperty, etc. Each Extracted concept is compared to pre-populated lists of characteristics. An example of a list of characteristics is as follows: HumanActionsList = [laugh, feel, love, talk, learn, think, read] or HumanDesiresList = [anger, feel, knowledge, happy, sad, surprised, afraid, envy, laughter].

```
Algorithm 2.   Recursive DFS feature extraction approach for
Sentient Nominal Entities
Input: T, BranchStack
Output: NP candidate with text-based features
Define Function ExtractEntities (T, BranchStack)
If  node ∈ T do
        If node = 'NP' and T.depth = 3 do
            MakeActorCandidateFeatures(T, BranchStack)
        oldtag ← nothing
        newtag ← nothing
        for child in T do
            if child has node do
                BranchStack.append(child.node)
                oldtag ← newtag
                newtag ← child.node
                EvaluateTagsSequenceSameLevel(oldtag, newtag)
            ExtractEntities (child,BranchStack)
        if BranchStack has value do
                BranchStack.pop()
```

Figure 3 Algorithm 2 – Sentence traversal

Special tags for semantic class annotations are also used to identify additional characteristics of the actors such as organization, location, and person. Pronoun relation features used heuristics to test for several conditions related to the presence of pronoun types in the current sentence (i.e. singular pronoun, plural pronoun, etc.) and the distance from the actor candidate to that pronoun. The distance from the pronoun to the actor candidate is calculated using a word distance metric. This calculation is only performed at the sentence level.

Phase 3: Speech Feature Extraction

Speech signal features can be incorporated in the sentient actor detection process to capture additional information about the speech content. Since emotions occur to sentient beings, it is conceivable to think that emotion related features could be correlated with the presence of actors in a section of an audio file. Many works such as Busso et al. (2009) and Luengo et al. (2010) argue that prosodic and spectral speech features such as pitch average, formants, and Mel Frequency Cepstral Coefficients (MFCCs) can help to capture some aspects of emotion in speech. Therefore, this work incorporates these features to see if they contribute to the task of sentient actor detection. The speech features used are those from the sentence where the actor appears. As such, each sentence where the NP candidate appears serves as a window of additional speech information. Speech features used in this work include pitch (F0), formants (F1-F5), and 12 MFCCs.

Phase 4: Classification

The methodology uses a total of 83 features from the different types described in the previous sections for classification purposes. Classification methodologies were implemented in WEKA and LibSVM, and included Support Vector Machines (Chang and Lin 2001), multilayer perceptron, random forests, and the nearest neighbor classifier.

Analysis and Results

In this section, the results of the training and testing methodology are presented. The sentient entity detection was performed on a dataset containing 4885 samples of which 2885 samples were labeled as non-actors and 2000 samples were labeled as actors. After performing feature extraction, the data was classified using several machine learning methods including Support Vector Machines, Artificial Neural Networks, Random Forests, and Nearest Neighbor classifier. These different methods are used to determine if the results of adding speech features are consistent across learning methods.

Table 1 Actor detection (for all actors) using text and speech features

Multiple Classifier Comparison (F-measure) **Training Set (3908) and Testing Set (977)**				
	SVM	**Multilayer Perceptron**	**Nearest Neighbor**	**Random Forests**
	Correct	Correct	Correct	Correct
Actor (Text)	79.1%	77.8%	75.6%	77.6%
Actor (Text & Speech)	81.8%	77.9%	80.1%	77.4%
Not Actor (Text)	87.4%	86.9%	83%	86.3%
Not Actor (Text & Speech)	88.9%	85.8%	87%	87.3%
All (Text)	84.2%	83.4%	80.1%	82.9%
All (Text & Speech)	86.1%	82.7%	84.3%	83.5%

The corpus is divided into two sub-sets; one for all entities (Tables 1 and 2) and one where named entities have been filtered out (Tables 3 and 4). The analysis of the F-measure accuracy scores for the first sub-set can be seen in Table 1. The results for the all entities analysis (Table 1) indicate that SVM performed best overall (86.1%) and that adding speech features improves actor detection accuracy.

Table 2 Results of feature analysis for the entire corpus

Rank	**Chi**	**Feature**
1	1569	Is a person? (Text: Knowledge)
2	758	NNP (Text)
3	740	Branch stack has no VP (Text)
4	503	Branch stack has no PP (Text)
5	220	Is female (Text: Knowledge)
6	199	SEM Org. (Text: Knowledge)
7	138	Human capabilities? (Text: Know.)
8	128	Human desires? (Text: Know.)
9	96	Is an animal? (Text: Knowledge)
10	79	Is a man? (Text: Knowledge)
11	78	NNS (Text)
12	70	Royal? (Text: Knowledge)
13	47	Is good? (Text: Knowledge)
14	45	MFCC 7 mean (Speech)
15	39	Branch has no SINV (Text)

The SVM classifier was trained using an RBF kernel with cost of 32 and gamma equal to 0.08. Results of the feature analysis (Table 2) indicate that semantic features derived from knowledge based approaches have a high contribution to sentient entity detection. Of the speech features, MFCCs performed best. To test the system's ability to detect unnamed entities (Nominal Entity Recognition), all named entities were filtered out. Once named entities were removed, the dataset consisted of 4019 samples of which 2735 were labeled as non-actors and 1284 samples were labeled as actors. Table 3 presents the results of the classification using the F-measure metric.

Table 3 Actor detection for unnamed entities

Multiple Classifier Comparison (F-measure) **Training Set (3215) and Testing Set (804)**				
	SVM	**Multilayer Perceptron**	**Nearest Neighbor**	**Random Forests**
	Correct	Correct	Correct	Correct
Actor (Text)	76.8%	72.6%	68.8%	76.1%
Actor (Text & Speech)	78.6%	75.9%	75%	71.2%
Not Actor (Text)	90.1%	88.8%	85.8%	90.2%
Not Actor (Text & Speech)	90.4%	88.7%	87.9%	89.1%
All (Text)	85.9%	83.7%	80.5%	85.8%
All (Text & Speech)	86.7%	84.7%	83.9%	83.5%

The results indicate that the SVM classifier performed best overall (86.7%) and that speech features help to improve classification accuracy. The SVM classifier used an RBF kernel with cost of 32 and gamma of 0.08. The results of the feature analysis (Table 4) indicate that a combination of knowledge-based, syntactic features, and speech features are useful for actor or nominal entity detection (unnamed entities). One important observation from the feature analysis is that the "Is a person" feature was consistently the highest ranked feature for all sets. This is important because it indicates that more knowledge can improve accuracy scores. For unnamed entity detection, it can be seen that speech features had a higher contribution to classification accuracy when compared to the previous feature analysis (for all entities) presented in Table 2. From the feature analysis (Table 4) it seems that most of the Mel Frequency Cepstral Coefficients (MFCCs) used in this work had some influence in the detection model. Finally, these results are consistent with recent results by Pang and Fan (2009a) and Pang and Fan (2009b) for Nominal Entity Recognition.

Table 4 Results of feature analysis for unnamed entities

Rank	Chi	Feature
1	864	Is a person? (Text: Knowledge)
2	504	Branch has no VP (Text)
3	394	Human capabilities (Text: Know.)
4	379	Female (Text: Knowledge)
5	357	Branch has no PP (Text)
6	287	Is m an (Text: Knowledge)
7	220	Hum an desires (Text: Knowledge)
8	189	Human properties (Text: Know.)
9	165	Is animal (Text: Knowledge)
10	162	Royal (Text: Knowledge)
11	91	Magical (Text: Knowledge)
12	75	MFCC 12 std. (Speech)
13	73	MFCC 7 mean (Speech)
14	61	MFCC 4 std. (Speech)
15	60	MFCC 5 mean (Speech)

Content Enrichment

Once sentient actors are detected in a speech file, the methodology automatically enriches the content by creating a new XML mark-up file with new metadata about the detected content. A scheme similar to the one used in Mallepudi et al. (2011) is implemented in this work. Figure 4 presents an example of the XML mark-up.

```
<?xml version="1.0" encoding="ANSI_X3.4-1968"?>
<audio format="MP3">
   <semantic>
      <Rule_inference>
          <type genre="fairy_tales"/>
           <type genre="children"/>
           ...
      </Rule_inference>
      <actors_List>
          <actor name="Tom" orderid="1" SentenceID="1">
             <characteristics magical="1" is_person="1" >
                 <sentence_position w="5"/>
          </actor>
           ...
      </actors_List>
           ...
</semantic>
</audio>
```

Figure 4 XML format enrichment for story in MP3 format

Conclusion

In this article, a methodology for actor detection and audio file enrichment is proposed. As can be seen from the results, the methodology obtained good results. Additionally, the results of the feature analysis on nominal entity detection indicate that speech features have an important contribution to the detection of these types of entities. After comparing multiple classifiers, the Support Vector Machines (SVM) model was found to be the best achieving an F-measure accuracy score of 86% for both named and unnamed entities. Additionally, feature analysis indicated that speech features were very useful when detecting unnamed actors.

Future work will focus on improving the methodology by incorporating anaphora resolution techniques so that the actors can be tied to all the referring expressions that make reference to them. Additional social metrics such as evolving emotional state of actors will also be analyzed and modeled.

References

Alm, C.O. 2008. Affect in Text and Speech, PhD Dissertation, University of Illinois at Urbana-Champaign, 2008.

Bigot, B., Ferrane, I., Pinquier, J., Andre-Obrecht, R. 2010. Speaker role recognition to help spontaneous conversational speech detection, In Proceedings of the 2010 international workshop on Searching spontaneous conversational speech (SSCS '10), ACM, New York, NY, USA, 5-10

Boersma, P. and Weenink, D. 2005. Praat: doing phonetics by computer, 2005, Retrieved from http://www.praat.org/.

Busso, C., Lee, S., Narayanan, S. 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, No. 4, May, 2009.

Calix, R., Knapp, G. 2011. Affect Corpus 2.0: An extension of a corpus for actor level emotion magnitude detection, In Proceedings of the 2nd ACM Multimedia Systems (MMSys) conference, Feb. 2011, San Jose, California, U.S.A.

Cassimatis, N. 2009. Flexible inference with structured knowledge through reasoned unification, IEEE Intelligent Systems, Volume 24, Issue 4, 2009.

Chang, C., Lin, C. 2001. LIBSVM: a library for support vector machines, Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chifu, E., Chifu, V. 2008. A Neural Model for Unsupervised Named Entity Classification, IEEE International Conferences on Computational Intelligence for Modeling, Control and Automation (CIMCA), pp. 1077-1082.

Coyne, B., Sproat, R. 2001. Inferring the environment in a text-to-scene conversion system, In Proceedings of Conference on Knowledge Capture, pp. 147-154.

Havasi, C., Speer, R., Alonso, J. 2007. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge, In proceedings of the 22nd Conference on Artificial Intelligence.

Johnson, J., Miller, A., Khan, L., Thuraisingham, B., Kantarcioglu, M. 2011. Extraction of expanded entity phrases, IEEE Intelligence and Security Informatics

Klein, D., Manning, C. 2002. Fast exact inference with a factored model for Natural Language Parsing, In Advances in Neural Information Processing Systems (NIPS 2002), Cambridge, MA: MIT Press, pp. 3-10.

LSU-NLP, Retrieve from http://nlp.lsu.edu/, 2011.

Luengo, E., Navas, Hernaez, I. 2010. Feature analysis and evaluation for automatic emotion identification in speech, IEEE Transactions on Multimedia, Vol. 12, No. 6, October 2010.

Mallepudi, S., Calix, R., Knapp, G. 2011. Material classification and automatic content enrichment of images using supervised learning and knowledge bases, In proceedings of the IS&T/SPIE International Conference on Multimedia Content Access: Algorithms and Systems V, January 2011, San Francisco, California.

McShane, M. 2009. Reference resolution challenges for an intelligent agent: the need for knowledge", IEEE Intelligent Systems.

Nguyen, H., Cao, T. 2008. Named entity disambiguation on an ontology enriched by Wikipedia, IEEE International Conference on Research, Innovation and Vision for the Future.

Pang, W., Fan, X. 2009, Chinese Nominal Entity Recognition with semantic role labeling, IEEE International Conference on Wireless Networks and Information Systems, 2009a.

Pang, W., Fan, X. 2009. A novel framework for Nominal Entity Recognition, Second International Symposium on Computational Intelligence Design, 2009b.

Stuker, S., Heck, M., Renner, K., Waibel, A. 2010. Spoken news queries over the world wide web, In Proceedings of the 2010 international workshop on Searching spontaneous conversational speech (SSCS '10). ACM, New York, NY, USA, 61-64.

Xi, Q., Li, F. 2011. Joint Learning of Named Entity Recognition and Relation Extraction, 2011 International Conference on Computer Science and Network Technology, Country, China, December 24-26