A Spoken Language Interface to an Electronic Programme Guide

Jianhong Jin[1]*, Martin Russell*[1]*, Michael Carey*[2]*,*
James Chapman[3]*, Harvey Lloyd-Thomas*[4] *and Graham Tattersall*[5]

[1]University of Birmingham, [2]University of Bristol
[3]BT Exact Technologies, [4]Ensigma Technologies, [5]Snape Signals Research
m.j.russell@bham.ac.uk

Abstract

This paper describes research into the development of personalised spoken language interfaces to an electronic programme guide A substantial data collection exercise has been conducted, resulting in a corpus of nearly 10,000 spoken queries to an electronic programme guide by a total of 64 subjects. A substantial part of the corpus comprises recordings of many queries from a small number of 'core' subjects to facilitate research into personalisation and the construction of user profiles. This spoken query data is supported by a second corpus which contains a record of subjects' viewing habits over a two year period. Finally, the two corpora have been combined to create two information retrieval test sets. Two probabilistic information retrieval systems are described, and the results obtained on the PUMA IR test sets using these systems are presented.

1. Introduction

PUMA is a collaborative project involving BT Exact Technologies, Ensigma Technologies, Snape Signals Research and The University of Birmingham. Its goal is to develop personalised user interfaces for portable or wearable devices, for improved human-machine interaction and user authentification. As an initial application, PUMA has focused on the development of a personalised spoken language interface to an Electronic TV Programme Guide (EPG). In this context, 'personalisation' means the construction of user profiles that encode a user's programme viewing preferences; and 'spoken language processing' involves retrieval of programme information in response to spoken queries.

One of the premises of the PUMA project is that at present, data driven statistical methods such as those employed in the AT&T 'How May I Help You?' system [1], are likely to provide the most successful approaches to robust spoken language processing. Such approaches typically attempt to derive a direct, probabilistic, relationship between the output of a speech recognition system in response to a spoken query, and semantic classes. However, while the number of classes in a system like HMIHY is small, in an EPG application there are many classes, corresponding to individual programmes. Hence there is unlikely to be sufficient training material to learn the required relationships entirely from data. The solution is to introduce an intermediate, and relatively low dimensional representation between spoken queries and semantic categories. In PUMA, two types of intermediate layer have been investigated. The first is entirely data-driven, based on Latent Semantic Analysis (LSA) [2], and the second uses explicit knowledge from WordNet [3]. This paper describes the second approach.

For either approach there is a need for data. A corpus of programme listings is needed to define the application, and, in the case of the LSA-based approach, to determine the intermediate layer. Information about the programme viewing preferences of individual subjects is needed to construct user profiles, and example spoken queries are needed to train and test the system.

The first part of this paper describes the various corpora which have been collected as part of the PUMA project. The second part presents a description of two of the PUMA information retrieval systems, and their performances on a large set of transcriptions of real spoken programme queries.

2. The PUMA Electronic Programme Guide corpora

2.1. The programme listings corpus

The programme listings corpus comprises a set of descriptions of all programmes shown on the five analogue British terrestrial TV channels (BBC1, BBC2, ITV, Channel 4 and Channel 5) after 6pm between September 2000 and November 2002. Each description includes the title, start and end times of the programme, plus a description of its content. Initially the listings were parsed from pages downloaded from http://sceneone.co.uk. However listings ceased to be available from this site in March 2002. Thereafter the listings were obtained from http:///www.bleb.org/tv.

2.2. The PUMA user profiles

Programme viewing profiles were compiled using data provided each day between September 2000 and November 2002. Each subject received a daily email, listing all programmes shown on UK terrestrial TV channels after 6pm on the previous day. Next to each title was a box, which the subject ticked to indicate that they had watched that programme. In all, 120 subjects participated in data collection and 41,015 programme viewings were recorded. For many subjects, participation in this exercise coincided with the start of recording of spoken EPG queries in June 2002, described in the next section. These subjects are of particular interest since data exists to create a user profile, from which prior probabilities of programme choice can be estimated, and to characterise their spoken queries. However, this type of user information was not used in the experiments reported here.

For example, at the University of Birmingham 17 'core' subjects returned daily questionnaires for an average of 8 months each. They watched 3,278 programmes over 138 'person months', giving an average monthly viewing rate of 23.75 programmes (less than one programme per day). These figures do not take into account days when subjects failed to return questionnaires, either as an oversight or because, for example, they

were on holiday. Hence the statistics are almost certainly an under-estimate of actual viewing frequencies.

2.3. The spoken queries corpus

A set of spontaneous spoken EPG queries was collected using a simple prototype real-time system. A spoken query was recognised using the SPRACH speech recognition system [5]. The resulting output was compared with programme listings for a seven day period using a simple word correlation algorithm. The top five programme listings were displayed and the subject was asked to indicate their relevance to the query. Although this system was rudimentary, it was sufficient for data collection purposes.
Spoken queries were collected at The University of Birmingham and BT Exact Technologies at Adastral Park. Sixty four subjects completed 468 sessions and recorded a total of 9,941 spoken EPG queries. Because of the emphasis of the PUMA project on personalisation and authentification, there was a need to collect a large number of queries from each subject. At the University of Birmingham, 18 subjects each completed an average of 15.17 sessions over a 20 week period and recorded an average of 348 spoken queries per subject. All but 1 of these subjects also completed the daily programme viewing questionnaire from the previous section. At BT Adastral Park, 44 subjects each completed an average of 4.43 sessions over 20 weeks and recorded an average of 84 spoken queries. These subjects will be used mainly as impostors in verification experiments. Summary statistics for the PUMA spoken queries corpus are shown in table 1.

Table 1: *PUMA spoken queries corpus statistics*

Site	University of Birmingham	BT Adastral Park	Overall
Num. of subjects	20	44	64
Tot. sessions	273	195	468
Tot. queries	6,264	3,677	9,941
Av. sessions per subject	13.65	4.43	7.31
Av. queries per subject	313	84	155

All of the spoken queries were transcribed at Birmingham. A set of XML tags was devised for annotation of repetitions and repairs.

2.4. The PUMA test set

A set of 400 spoken queries was chosen randomly from the 6,264 queries spoken by the Birmingham subjects during the period May 2002 to September 2002. All queries, and those listings from the evaluation period which corresponded to a programme shown between the 20th and 30th day of the month, were tagged manually with one or more broad programme categories. There were 20 such categories, for example 'sport' and 'documentary'. These tags were used as a filter to facilitate the manual association of a set of possible queries with each programme listing. Finally, for each query the set of programme listings associated with that query in the previous stage were compiled to form the set of 'correct' listings for the query.
This set was split into two halves. One half was intended to be used for automatic learning of relationships between query words and words in the corresponding programme listings, while the other was used as the test set. In fact, it was not possible to associate any listings with 15 of the queries in the test set. These queries were discarded, leaving a 'core' PUMA information retrieval test set of 185 queries.
An interesting issue is the correct treatment of 'series' programmes. For example, should different episodes of 'The Simpsons' be treated as programmes in their own right or combined into a single 'meta-programme'? An additional version of the test set was created in which the listings for series' of recurring programmes were combined into a single listing. This new test set and the original test set are referred to as the *contracted* and *expanded* test sets, respectively.
The expanded test set contains 2,371 programme listings with a mean document (listing) length (the number of words in a programme listing and its title previous to any pre-processing) of 32. For the contracted test set the number of programme listings is 870, and the mean document length is 113.
Although outputs from the SPRACH speech recogniser were available for all of the queries, the manual transcritions of the queries were used in the experiments reported in this paper.

3. Information retrieval systems

Two retrieval systems were evaluated on the TV listings data. Both are based on the probabilistic approaches described in [6], and use the thesaurus dictionary WordNet [3].
The text corresponding to a spoken query or programme listing in the evaluation set is pre-processed as follows: First all of the stop words are removed and the remaining words stemmed. Then the stemmed words are replaced with symbols representing semantic classes which they belong to. Finally index files are generated which specify the relationships between all the possible terms and programme listings in the evaluation database. Two types of semantic class were considered: word stems and WordNet synonym classes.
In the results section, the information retrieval system based on word stems (or word form) is referred to as 'WF', while the one based on WordNet synonym classes is referred to by the acronym 'WN'.

3.1. Stop words

'Stop words' are non-content words which are judged not to contribute to the information retrieval process and are thus treated as noise. These are removed to make the system more efficient. One commonly used approach to building a stop list is to use one of the many lists which have been generated in the past for previous applications. A more suitable method is to produce a word frequency listing for the text in the current application, and then to examine each of the high frequency words. If there is no known importance of a given word in the application, then that word can be safely placed on a stop list [7]. However, sometimes high frequency words are important, and additional criteria are needed to identify stop words reliably. At the moment, only a very short stop word list is used in our systems.

3.2. Stemming

Stemming is used to identify different forms of the same word, so that these different instantiations of the word can be treated appropriately for the purposes of information retrieval. The WordNet function 'morphword' is used to convert words into their base forms. Two lists are used for this function. These

are lists of inflectional endings, based on syntactic category, that can be detached from individual words (for example, word "ponies" to "pony"). There are also exception list files, one for each syntactic category, in which a search for an inflected form is done. The exception lists contain the morphological transformations for strings that are not regular and therefore cannot be processed in an algorithmic manner. In the case of the WF retrieval system, no further categorisation of words is used. However, for the WN system a further stage is employed in which stemmed words are allocated to synonym classes.

3.3. Synonym groups

Words are classified into synonym groups by consulting WordNet. Each word belongs to at least one synonym group, and words will belong to several if they have multiple senses. In the case of 'out of vocabulary' words (i.e. words which do not occur in the WordNet dictionary), it is not possible to find any meaning from WordNet, and so it is not possible to find a synonym group. As one might expect, many such words occur in TV programme descriptions. In such cases a new synonym group was defined, which contained only this word. According to this strict definition of synonym, programme listings which contain words which do not have exactly the same meaning, but only have related meanings, to words in a query will be rejected. Hence additional WordNet word relationships, including Hypernym, Hyponym, Meronym, Holonym are used to establish relations between synonym groups.

3.4. Index files

The index files identify, and contain information about, those word classes which are useful for discriminating between the different TV programme listings in the database. They are generated during a 'pre-processing' stage and are used by the retrieval system to speed up the process of relating words in a query with TV programme listings, thus obviating the need to search through each document in real time. Each line in an index file starts with a *key* representing a synonym group, followed by the number of programme descriptions that include the key, and their identities.

3.5. Decision making

The following function ([7], [8]) computes scores for each word in a query ($w_i, i = 1, ..., N_q$) with respect to each document ($d_j, j = 1, ..., N$). Note that the word may be represented as a group of related keys if synonym group expansion has been used, as in the WN system.

$$cw(w_i, d_j) = \frac{CFW(w_i) \times KF(w_i, d_j) \times (a+1)}{a((1-b) + b(NDL(d_j))) + KF(w_i, d_j)} \quad (1)$$

where $CFW(w) = log(N) - log(n_w)$. Here N is the number of listings in the evaluation set and n_w is the number of listings which include any key related to w, $KF(w_i, d_j)$ is the number of occurrences of all keys in the expanded group for word w_i in document d_j.

$$NDL(d_j) = \frac{DL(d_j)}{DL_{av}}, \quad (2)$$

where $DL(d_j)$ is the total number of key occurrences in document d_j and DL_{av} is the average number of keys per document. The parameters a and b are tuning constants, which modify the influence of term frequency, and document length, respectively. When calculating KF, the synonym keys in the key-group should be given higher priority than other semantic types.

$$KF(k_i, d_j) = \sum_t h_t P_t(k_i, d_j) \quad (3)$$

where t represent the type of semantic relationship group that K_i belongs to. $P_t(k_i, d_j)$ is the number of occurrences of synonyms, hyponym, hypernyms, meronym, holonym of key k_i in document d_j, respectively. h_t ($t = 1$ to 5) are tuning parameters to adjust the importance of synonyms and other semantically related groups of a key. Synonyms are considered more important than hyponyms and hypernyms, so h_1 is typically given a larger value than the others.

4. Experiment and Results

Recall that the manual transcriptions of the spoken queries in the test set, rather than the outputs of the SPRACH recogniser, were used in all experiments.

Figure 1 shows precision-recall graphs for the two systems (WF based on stemming, and WN based on WordNet synonym groups) on the expanded test set (i.e. on the test set which contains separate listings for different episodes of the 'same' programme). The graphs indicate that the performances of the two systems are very similar on this test set, with the system based on wordnet synonym groups giving the best performance. Note also that the mean average precision scores achieved by the two systems (0.49338 and 0.50827, respectively) are comparable with those achieved on the DARPA Spoken Data Retrieval (SDR) task [9].

The results for the contracted test set (in which listings for different episodes of the 'same' programme are combined into a single listing) are shown in figure 2. In this case, the performances of the two systems are even closer. Overall, the perfomances on the contracted set are better than that on the expanded set, because the programme listings in the contracted set contain more words (the average is 113 for the contracted set and 32 for the expanded set) and there are fewer documents (870 instead of 2,371). The fact that the MAP scores for the WF and WN systems are closer for the contracted set may be because the longer listings already contain many of the relevant synonyms.

5. Conclusions

This paper is in two parts. The first part describes the substantial data collection exercise that has been conducted as part of the PUMA project. This has resulted in a large corpus of nearly 10,000 spoken queries to an electronic programme guide, spoken by a total of 64 subjects. In order to support personalisation and the development of user profiles, a substantial part of the corpus comprises recordings of many queries from a small number of 'core' subjects, collected over a 20 week period. This spoken query corpus is supported by a second corpus which contains a record of subjects' viewing habits over a two year period. Each of the core subjects from the spoken queries corpus is also included in this second corpus. Finally, the two data sets have been combined to create two information retrieval test corpora. Both use the same set of transcriptions of 185 spoken queries. In the first (expanded) test set, these queries are related to a set of 2,371 programme listings, while in the second (contracted) test set, listings corresponding to different episodes of

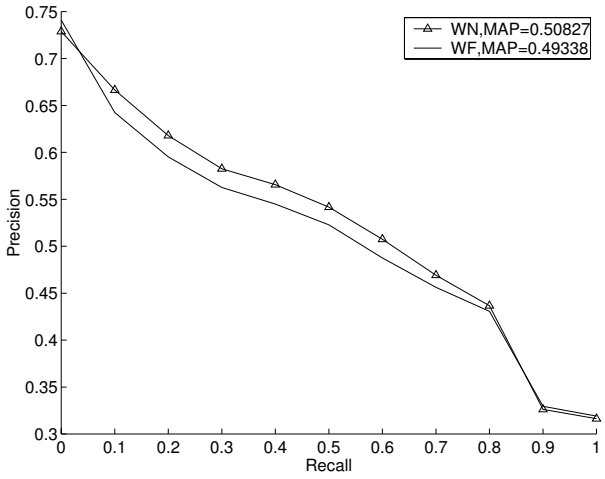

Figure 1: Precision-recall curves for the expanded test set using the retrieval systems based on word stems (WF)(−) and WordNet synonym classes (WN) (△)

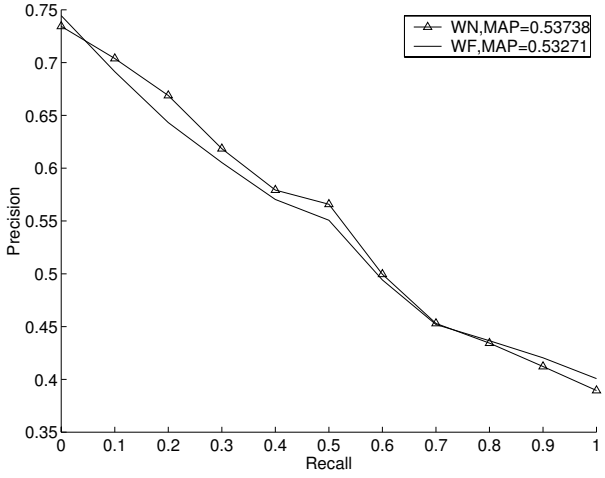

Figure 2: Precision-recall curves for the contracted test set using the retrieval systems based on word stems (WF) (−) and WordNet synonym classes (WN) (△)

the same programme are combined, resulting in 870 listings in total.

In the second part of the paper, two probabilistic information retrieval systems are described, one using word classes based on stemming (referred to as system WF), and the other using WordNet synonym classes (referred to as WN). The results obtained on the PUMA IR test sets using both of these systems are presented. For the expanded test set, both systems achieve similar performance, with mean average precision scores of 0.50827 (WN) and 0.49338 (WF). The perfomance of both systems (mean average precision 0.53738 for WN, 0.53271 for WF based system) on the contracted set are superior to that on the expanded dataset, which may be caused by longer descriptions and smaller number of programme listings. Also, the MAP values for the two systems are closer on the contracted set. This is attributed to the fact that many relevant synonyms will occur in the combined listings, obviating the need for WordNet synonym classes in this case.

6. References

[1] Gorin, A.L., Abella, A., Alonso, T., Riccardi, G. and Wright, J.H., "Automated natural spoken dialogue", IEEE Computer Magazine, vol 35(4), pp. 51-56, 2002.

[2] Landauer, T.K. and Dumais, S.T., "A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge", Psychological Review, 104, 211-240, 1997.

[3] Miller G.A., Beckwith R., Fellbaum C. Gross D., and Miller K., "Introduction to WordNet: An On-line Lexical Database", http://www.cogsci.princeton.edu/ wn, August 1993.

[5] Cook, G., Christie, J., Ellis, D., Fosler-Lussier, E., Gotoh, Y., Kingsbury, B. Morgan, N., Renals, S., Robinson, T. and Williams, G., "A overview of the SPRACH system for the transcription of Broadcast News", Proceedings of the DARPA Broadcast News Workshop, February/March 1999.

[6] Robertson, SE, and Sprk Jones K. (1997), "Simple proven approaches to text retrieval", Tech. Rep. TR356, Cambridge University Computer Laboratory, 1997.

[7] Harman, D., "NIST Interagency Report 4873: Automatic Indexing", National Institute of Standards and Technology, Chapter 1, p3, July 1992.

[8] Spark Jones K., Walker S. and Robertson S. E., "A Probabilistic model of information retrieval: development and status", TR 446, Cambridge University Computer Laboratory, September 1998.

[9] Voorhees, E.M., Harman, D., "Overview of the seventh text retrieval conference (TREC-7)", Proceedings of the Seventh Text REtrieval Conference (TREC-7) held in Gaithersburg, Maryland, USA, 1998.