Heuristic Selection of Training Sentences from Historical TV Guide for Semi-supervised LM Adaptation

Harry M Chang

AT&T Labs – Research, Austin, TX 78795, USA

Harry_chang@labs.att.com

Abstract

This paper describes a novel approach to the automatic selection of training sentences from a system-generated data feed for the development of high-precision language models (LMs) required for speech-enabled voice interface applications in the TV search domain. We develop a set of heuristic rules to select training sentences directly from the TV *electronic programming guide* (EPG) in their metadata form. The training corpus constructed using the selection algorithms encoded with the historical EPG data enables the adapted LMs to have a considerably lower perplexity while achieving a significant reduction in word error rate (WER). When evaluated using the user-generated spoken queries to an experimental TV search application, a 10% absolute reduction of WER is reported over the baseline LMs created without using the training sentences generated from the historical EPG data.

Index Terms: spoken language modeling, speech recognition, text mining, TV electronic programming guide

10.21437/Interspeech.2013-524

1. Introduction

There is a growing interest in the development of speech-driven interface technologies for a common entertainment activity at home: television viewing [1, 2, 3, 4]. Creating a speech-driven search interface for TV has some unique challenges in terms of building high-precision language models (LMs) for automatic speech recognition (ASR) technologies. There are two major obstacles facing the developers when relying on the search term coverage from traditional n-gram based LMs [5, 6, 7] for such a large-vocabulary ASR application. First, the underlying content in a TV guide changes daily; therefore, the user language models will alter with the ever-shifting content on TV. Secondly, it is often impractical to collect an adequate training corpus from user-generated queries in spoken format before it becomes obsolete.

2. Constructing training corpus from EPG

The primary motivation for acquiring the training sentences directly from the EPG metadata source is driven by two practical considerations for any major TV service provider such as AT&T: a) daily availability of the XML-based EPG metadata source covering all scheduled TV programs for the next 15 days, and b) the richness of the text content embedded in such a data feed. Our main objective is to build a training corpus directly from the system-generated EPG metadata source and then use the corpus to create the LMs for an ASR engine. The criterion for the LMs is to obtain a sufficient and efficient coverage for those queries most likely to be spoken by average TV viewers at home, through applying domain heuristics to achieve a higher recall and lower perplexity.

2.1. EPG Ontology and Data Structures

EPG metadata sources for a TV guide used by the TV industry in the U.S. are typically structured in an XML format. The EPG metadata can have as many as 75 distinct types of data elements. Our previous study [8] shows that even a simplified taxonomy for a commercial EPG data feed can have over 25 content nodes. Figure 1 provides a flattened view of a single data record instance from a highly simplified EPG ontology tree. In this example, only 14 data elements (printed in italic) are explicitly defined (e.g., *title*, *actors*, *genres*, etc.).

<table>
<tr><td colspan="3">*Title*: “The World is Not Enough”</td><td>*Type*: Movie</td></tr>
<tr><td colspan="2">*Director*: Michael Apted</td><td>*Rating*: PG-13</td><td>*Year*: 1999</td></tr>
<tr><td colspan="4">*Actors*: Pierce Brosnan, Robert Carlyle, Judi Dench, …</td></tr>
<tr><td colspan="2">*Genres*: Action, Adventure</td><td colspan="2">*Country*: USA</td></tr>
<tr><td rowspan="2">*Show Schedule*</td><td>*Channel Callsign*: “HBO”</td><td colspan="2">*Channel#:* 802</td></tr>
<tr><td>*Time*: 2011-11-17 19:00”</td><td colspan="2">*Duration*: 128m</td></tr>
<tr><td colspan="4">*Program Description*: James Bond goes to bat for queen and country by protecting an oil heiress from the terrorist who killed her father and retrieving M from his evil ….</td></tr>
</table>

Figure 1: *A data record instance of EPG*

2.2. Create EPG ontology using domain heuristics

In our previous studies [9, 10], we analyzed the user-generated search terms (both *spoken* and *typed*) in a controlled laboratory environment and from a web-based EPG application. We discover that a very small set of EPG data categories, such as *program titles,* have an overwhelming influence on the user’s language model. A simple statistical analysis of word frequency reveals that the *title* words in a typical EPG data source accumulated over a 12-month period represent over 95% of all word instances in the underlying TV guide. The empirical data analyzed during our studies suggest that over 93% of all user-generated search terms can be predicted with the texts associated with only 7 data categories in the EPG metadata. For this paper, the heuristic selection of training sentences is limited to the texts in these 7 EPG sub-categories as listed below.

PT: Program Titles (“`Seinfeld`”, “`America Idol`”)

CD: Channel Descriptions (“`CNN`”, “`Disney HD`”, “`HBO`”)

CN: Channel Numbers (“`302`”, “`1602`”, “`Channel 5`”)

MA: Movie Actors (“`Tom Hanks`”, “`Eva Green`″)

MD: Movie Directors (“`Woody Allen`”)

TC: TV Cast (“`Tina Fey`”, “`Beyonce`”)

SC: Sports Cast (“`Dallas Cowboys`”, “`David Beckham`”)

2.3. Sentence Generation from the EPG metadata

Our prior study [10] shows that the total number of unique sentences in the PT category over a 2-year period can exceed 90,000. Even with a constrained *n*-gram generation process

[11], a training corpus from title sentences *alone* could contain over a million *n*-gram phrases ($n \leq 4$). To reduce the over-generation of *n*-grams from title texts that are unlikely to occur in the user-generated spoken queries, we impose more restrictions on sentence generation based on a 50-year-old theory known as the principle of least effort [12, 13]. The restrictions on the single pass left-to-right parser [11] are introduced to mimic the users' mental model based on a small set of heuristic rules for this domain. The principal assumption is that average TV viewers tend to compose spoken queries with the least number of words associated with a specific TV title or genre as possible, without self-induced ambiguity.

2.3.1. Selecting Title Sentences

The heuristic rules implemented for this study do not require an in-depth knowledge of the underlying semantic structure of the texts to be parsed. Only true substrings of a raw sentence are allowed with a conditional cap: the length of all training sentences selected ≤ 7 words. A list of stop words for common propositions (e.g., "*and*", "*in*", "*with*", etc.) are created to enable the parser to remove those instances only from the phrase initial and final. For example, for a 7-word program title "`All Bets Are off with Bruce Drennan`", the 5-word phrase candidate "`All Bets Are off with`" is generated first. Based on the heuristic rules, the phrase ending word "*with*" must be omitted. The edited phrase is then used as a training sentence. This word omission rule can be overridden by another rule: all title sentences with 5 words or less are retained. Altogether, approximately 250,000 unique title expressions are constructed from the EPG text corpus created from the metadata feed over a 4-year period. For the PT category, the average sentence length is 3.1 words.

Each title sentence *s* in the PT data set is assigned a rating score, *p(s),* primarily based on the accumulative schedule frequency *f* of the corresponding TV show on the broadcasting channels over a given time period *t* as defined by (1). The constant *k* with an integer value between 1 and 6 is used to introduce a positive bias for programs broadcasted during *primetime* and for programs that are *new* (have never before been broadcasted).

$$p(s) = \sum_{t}^{n} f(s) * k \mid s \in PT \tag{1}$$

2.3.2. Selecting Names from Cast Categories

The text strings in a *cast* data field are treated as a baseline name entry such as "`Samuel L. Jackson`". A rule-based transformation is applied to the baseline entry to produce one or more name expressions. For example, the name initial '`L.`' in a name string would be skipped if it is followed by a text string with at least 2 characters. There are approximately 150,000 baseline entries in the combined cast categories. As a result of the transformation, the union of all four cast categories (*MA*, *TC*, *SC*, and *MD*) contains over 250,000 name expressions.

To account for the greater influence of the *leading* cast members on a program title in terms of the likelihood that users may include their name in a spoken query, the first *three* cast members (typical program titles have 6 cast members) are given a progressively higher credit score *c* (from 6 to 4) for each appearance in a TV program. All other cast members receive one credit score for each appearance. The overall rating score for each name expression *s* in all four cast categories is computed using (2), based on the accumulative frequency *f* of the shows where *s* is derived from a cast member listed under the show title. Similarly to (1), a bias factor *k* is used to increase the weights for those program titles shown during *primetime* and those listed as a *new* episode.

$$p(s) = \sum_{t}^{n} f(c_s) * k \mid s \in cast \tag{2}$$

From the PT data set, a subset of movie titles (MT) is created if the *type* of program title in the EPG metadata is defined as *movie*. From the combined cast category, three subsets MA, TC, and MD are created based on the *type* of programs (*movie* or *TV series*) and the *cast* sub-categories: *actors* or *directors*. Finally, cast members associated with any program titles that have ever been broadcasted on 6 well-known sports channels such as ESPN are grouped together to form the SC data set. For each training sentence with dynamic content in these 6 EPG categories (PT, MT, MA, MD, TC, and SC), their overall rating score *p* is dependent upon the length of the rating windows of last *n* days from a starting date *t*.

2.4. Determine Length of Rating Windows

Generally speaking, the longer the rating windows, the higher the perplexity of the resultant LMs, since more training sentences will have been added. Heuristically, we have a simple hypothesis: different EPG categories may require different rating windows. For example, average TV viewers may have a relatively longer memory for some hit movies than for those TV series lasting only a few seasons. Similarly, sports fans may remember their favorite teams for far longer than for the names of a TV celebrity whose fame may only last a few months. To study the accuracy effect of the LMs created from the training sentences selected from different rating windows, we revise the common *k*-fold cross-validation technique [14] to allow the time overlap while focusing on a single learning parameter: the length of a rating window. The entire historical EPG guide is partitioned into *four* overlapping rating windows: previous 6 months, 12 months, 24 months, and 48 months from a given starting date *t*.

First, a baseline LM is created only using the sentences selected from a 15-day EPG that aligns with the same time period when the test utterances were collected. Secondly, 24 new LMs are created by including additional training sentences selected from 6 EPG dynamic content categories over a longer rating window. In all ASR tests, a standard W3C SRGS grammar construct is used to create the LMs for the AT&T WATSON(SM) speech recognizer [15]. A small pilot set, P1K, consisting of the first 1000 utterances in the test set Nov2011 (Section 3.1) is used to derive the *relative* performance *delta* as reported in Table 1. The symbols "+" and "-" in Table 1 indicate if the expanded LMs outperform or underperform the baseline model in relative percentages.

Table 1. *Accuracy effect of rating windows*

EPG Categories	-6m	-12m	-24m	-48m
TV Titles	+7	+8	-2	-3
Movie Titles	+4	+5	+3	-2
Sports Cast	+2	+2	+2	-1
TV Cast	+2	+3	-2	-3
Movie Actors	+3	+3	+3	-2
Movie Directors	+1	+2	+1	+1

As a result of this simple data-driven learning process, the duration of the rating windows for the 6 categories are determined for the subsequent process: 12 months for TV Titles, 24 months for Movie Titles, etc. The lengths of these rating windows are consistent with our observations from analyzing the user-generated queries for our prior studies [8, 9, 10]. For example, for average TV viewers, the name recall for a TV cast tends to fade away after 12 months if the cast member is no longer appearing in any TV show since then.

2.5. Reducing Perplexity of LMs

The expanded LMs has a very high recall (>95%) and it easily outperforms the baseline LMs in terms of WER. But the absolute WER figure is too high (>40%) for the targeted ASR applications envisioned. A simple analysis reveals that most of the substitution errors can be traced to either TV titles with a very low rating score or to the names of relatively unknown movie actors. Excluding these training sentences from the expanded LMs would considerably reduce the perplexity of the final LMs.

Within each of the 6 dynamic content categories, all training sentences selected have a uniform rating scale as computed using (1) and (2). Therefore, a single threshold can be used to exclude training sentences with relatively lower rating scores. However, the optimal cut-off threshold is difficult to derive without introducing supervised learning, which needs a large amount of user-generated queries for the domain. For the sake of simplicity, we decide on a relative threshold by excluding sentences with rank scores in the bottom 10 percent of their respective data categories. This method turns out to be highly effective. For example, the total number of bigrams in the new PT training set is reduced by 26% for a typical 15-day rating window.

According to the web-based user queries for TV search we collected from our previous studies [10,11], we believe that the total number of cast names is still too high, even after applying the exclusion threshold. Heuristically, we would predict that the same user population would follow the same search pattern, regardless of input modality (typing vs. speaking). Following this logic, we further reduce the size of the TC and SC training sets by including only those with above a certain rank. For TC, 1200 names with the highest rank scores over a 12-month rating window are chosen for the expanded LMs. Similarly, for SC, 400 names with the highest rank scores over a 24-month rating window are selected for the expanded LMs for the subsequent ASR tests.

All training sentences selected for their corresponding final SRGS sub-grammars have a uniform weight within each SRGS grammar rule scope. Different grammar rules are given a different category weight based on the domain knowledge we learn from our previous studies [8, 9, 10]. For example, the *title* search category is given the highest weight in the SRGS grammar. This is because the web-based user queries are dominated by the phrases intended for a program title such as "*Seinfeld*", "*The Big Bang Theory*", "*The Voice*", etc.

3. Performance Evaluation

The main objective of this study is to automatically generate training sentences by applying a set of domain-driven heuristic rules to the text selections from the EPG data feed accumulated over time. The main criteria used to evaluate the quality of the resultant training corpus consists of a) WER on the resultant LMs created from the incorporation of the historical EPG data and b) measurement of the perplexity of the LMs. Unlike common web-based text search applications, for voice search apps, it is rather easy to achieve a high recall by simply adding more and more training sentences to the LMs in order to cover 90% of all potential search terms. But the over-generation of training sentences always leads to a higher WER.

3.1. Test Set – Speech Utterances for TV Search

The two sets of speech utterances are collected using an iPhone-based EPG test application developed by the author. The data collection app uses the actual EPG metadata feed for the U-verse TV service provided by AT&T, and provides real-time feedback for the spoken search queries. This provided a realistic simulation for the speech-enabled voice search applications envisioned, where targeted users would use this app in front of the TV in their own homes. The first set contains 1,394 utterances collected over a 4-week period in November 2011 and is referred to as the Nov2011 data set, where most of spoken queries were recorded after November 17. The second set, containing 2,100 utterances, was collected over a 3-week time period in March 2012 and is referred to as the Mar2012 data set, where most of spoken queries were recorded before March 10. Both data sets came from a relatively small user population (~100). The data collection participants are company employees, and most of them are located in Austin, Texas

3.2. Speech Recognition Test Configurations

Since the test sets (Nov2011 and Mar2012) were collected from two different time periods, it is expected that the users' search terms are also different. Therefore, two different anchoring dates (t_0) for all rating windows are established, one on 11/17/2011 for the Nov2011 test set and the other on 2/24/2012 for the Mar2012 test set. The both anchor dates are selected in such a way to overlap the data collection periods.

In addition to the training sentences selected from the 6 EPG categories with dynamic content, training sentences for other 2 categories (CD and CN) with relatively static content as listed in Table 2 are generated from the EPG metadata. The sentence selection for the CD data set is based on the domain heuristics learned from analyzing the web-based user expressions related to a channel name in terms of phrase variations (e.g., "*HBO*", "*Home Box Office*", "*HBO HD*", "*HBO High Definition*", etc.).

To increase the coverage for older movies and/or TV shows, we also create a list of the top-100 TV shows over the last 40 years and a short list of the top-rated movies based on information provided from a few widely-visited websites. The title expressions from these web-harvested lists are added to the training corpora for the baseline LMs. Finally, a list of 50 short phrases for TV/DVR control (e.g., "*Channel Up*", or "*Fast Forward*") is added to the training corpora.

Table 2. *Training sentences for the baseline LMs*

EPG categories with relative static content	# of Phrases	Examples
CD: Channel Description	~900	"*Disney Channel*"
CN: Channel Number	~600	"*Channel 302*"
Top 100 TV shows	170	"*I Love Lucy*"
Top 50 Movies	55	"*Godfather*"
TV/DVR Control	50	"*Channel Up*"

3.2.1. *Training corpus for the baseline LMs*

To cover the potential search phrases spoken by the data collection participants, a training corpus is constructed for each of the two baseline LMs, one to be used for the Nov2011 test set and the other for the Mar2012 test set. The same exclusion threshold described in Section 2.5 is used to remove any sentence with a rank score in the bottom 10th percentile in their corresponding category. There is one exception - all program titles marked as a *new TV show* in the corresponding 15-day TV guide are included in the training set regardless of their ranking score. Table 3 summarizes the training sentences generated from the 6 dynamic categories in the 15-day EPG metadata for the two baseline LMs.

In anticipation of the possible framing phrases average TV viewers may use to express their search intention, we select 10 simple framing phrases for various search categories and incorporate them in the form of prefix or postfix in various slots in the SRGS grammars files. Two examples of this type of framing phrase for the <*movieActor*> slot are "`movies with` <*movieActor*>" and "<*movieActor*> `movies`". Similarly, we added a few such framing phrases for the <*programTitle*> slot with a prefix such as "`find`" and "`find the new episodes of`".

Table 3. *Training sentences for the baseline LMs*

EPG Categories	Nov2011	Mar2012
PT|*all*	15day: 7,767	15day: 7,724
MA	15day: 10,736	15day: 11,147
TC+SC	15day: 7,967	15day: 8,186

3.2.2. *Training corpus for the adapted LMs*

For the two benchmark tests, their corresponding baseline LMs are adapted by adding and/or merging the new sentences selected from the 6 dynamic content categories over the longer rating windows. If a new sentence created from the historical rating window is already in the baseline training corpus, no extra weight is given.

Table 4. *New training sentences for the adapted LMs*

EPG Categories	Nov2011	Mar2012
PT|non-movie	12mo: 20,464	12mo: 21,017
PT|movie	24mo: 1,200	24mo: 1,200
MA	24mo: 15,000	24mo: 15,000
MD	48mo: 100	48mo: 100
TC	12mo: 1,200	12mo: 1,200
SC	24mo: 400	24mo: 400

3.3. Test Results

Four LMs are created using the identical W3C SRGS grammar construct to organize the training sentences selected for the two ASR benchmark tests. For each benchmark test, the same AT&T WATSON$^{(SM)}$ speech recognizer is configured to run the two file-based batch tests, the first using the baseline LMs and the second using its adapted LMs as described in Section 3.2. The WER figures for the four tests are reported in Table 5. Word errors due to Out-of-Vocabulary (OOV) are extremely low (<0.5%) and therefore it is not reported here.

Table 5. *Word error stat from the ASR tests*

Test Sets	Nov2011	Mar2012
Baseline LMs	28.2%	33.3%
Adapted LMs	18.7%	19.7%

3.4. Discussions

The model coverage (recall measurement) from the training corpora created using the rating models described in this paper is over 98%, comparable with most commonly used *n*-gram based statistical language models (SLM) when trained using the same set of the sentences. Of all misrecognition errors from both benchmark tests, less than 3 percent can be identified as out-of-grammar, such as "*Smash on NBC*". This is because the training sentences constructed using our method are derived from a single content category such as title set (*PT*) or channel names (*CD*), but not both. In this particular error case, both the program title "*Smash*" and the channel name "*NBC*" are in the training set as a stand-alone phrase, but are not selected together to form a new training phrase. Other prominent errors are due to the lack of framing words before a content phrase, such as "*I want American Idol*", where the first two words are not part of any program title but are used by the speakers to *frame* their primary intention for a popular tv show called "`American Idol`" in the underlying TV guide.

The other error category is mainly caused by word substitutions and/or insertions. In those cases, the user either *replaces* a word in a program title with other words based on their memory recall or *transpose* around the actual words. A few errors are due to a lack of coverage for the *program description* field in the EPG metadata field such as "*James Bond's movies*". We purposely exclude this particular data category from the models because it increases the model perplexity by 5-fold.

4. Conclusion & Future Work

This paper describes a knowledge-based rating system capable of selecting high-value sentences from the historical TV guide in an unsupervised fashion. The adapted LMs outperform the baseline LMs by a significant margin in terms of precision measurement, while still maintaining a sufficient recall rate. The rating algorithms proposed in this paper can be implemented and run on the daily EPG data feed without any manual supervision. The optimal lengths of the rating windows for different EPG content categories can be systematically derived by using a small set of transcribed utterances from actual users after such an application is deployed.

Future research is required in order to expand the existing domain taxonomy to include the *program description* category. This would dramatically increase the perplexity of the LMs by a 5-fold if all *n*-gram substrings are included. New rating algorithms are needed to determine why some phrases (e.g., "*James Bond*") in the *program description* (PD) field deserves a selection and a much high rank score than other bigram phrases in the same data set (e.g., "*oil heiress*") for the movie example given in Section 1. We also plan to explore various shallow semantic parsing techniques to automatically select high-value terms in the EPG data feed, which will further expand the model coverage.

5. References

[1] Writtenburg, K., Lanning, T., Schwenke, D., Shubin, H., Vetro, A., "The Prospects for Unrestricted Speech Input for TV Content Search", In Proceedings of the Working Conference on Advanced Visual Interfaces, 352-359, 2006.

[2] Johnston, M., D'Haro, L-F., Levin, M., Renger, R., "A multimodal interface for access to content in home", ACL, 376-383, 2007.

[3] Renger, B., Feng, J., Dan, O., Chang, H., Barbosa, L.: "VoisTV: Voice-Enabled Social TV", WWW Companion Volume, 253-256, ACM, 2011.

[4] Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R.: "Analysis of the Characteristics of Talk-show TV Programs", InterSpeech 2012.

[5] Yamamoto, H., Isogai S., Sagisaka, Y.: "Multi-class Composite N-gram Language Model", Speech Communication, vol. 41 Issue(2-3):369-379, 2003.

[6] Roark, B., Saraclar, M., Collins, M.: "Discriminative N-gram Language Modeling," Computer Speech and Language, 21(2): 372-392, 2007.

[7] Gillot, C., Cerisara, C., Langlois, D., Hanton, J-P.: "Similar N-Gram Language Model", InterSpeech, 1824-1827, 2010.

[8] Chang, H.M., "Constructing n-gram rules for natural language models through exploring the limitations of the Zipf-Mandelbrot Law", Computing, vol. 91:241-264, 2011.

[9] Chang, H.M., "Conceptual modeling of online entertainment guide for natural language interface", 15th International Conference on Applications of Natural Language to Information Systems, 253-253, 2010.

[10] Chang, H.M., "Topic interference by weighted mutual information measures computed from structured corpus", 16th International Conference on Applications of Natural Language to Information Systems, 64-75, 2011.

[11] Chang, H., "Enriching domain-specific language models using domain independent WWW n-gram corpus", In Artificial Intelligence and Soft Computing, 38-46, Springer Berlin/Heidelberg, 2012.

[12] Zipf, G.K., "Human Behavior and Principle of Least Effort", Addison-Wesley, 1949.

[13] Vogt, P.I., "Minimum Cost and the Emergence of the Zipf-Mandelbrot Law", Proc. of the 9th Artificial Life Conference, MIT Press, Cambridge, Mass. USA, 2004.

[14] Dietterich, T.G.: "Approximate Statistical Test for Computing Supervised Classification Learning Algorithms", Neural Computation, vol. 10(7): 1895-1923, 1998.

[15] Goffin, V., Allauzen, C., Bocchieri, E., Hakkani-Tur, D., Ljolje, A., Parthasarathy, S.: "AT&T WATSON Speech Recognizer", ICASSP 2005.