SAME NEWS IS GOOD NEWS: AUTOMATICALLY COLLECTING REOCCURING RADIO NEWS STORIES

Stefan Rapp[*] *& Grzegorz Dogil*

Institut für Maschinelle Sprachverarbeitung (IMS),
Universität Stuttgart, Azenbergstr. 12, 70174 Stuttgart, Germany
e–mail: rapp@sony.de, dogil@ims.uni-stuttgart.de

ABSTRACT

We present methods for finding same or almost same news stories in the hourly radio news broadcasts. Our procedures are able to detect reoccuring news stories of subsequent news broadcasts spoken by the same or different announcers only from the speech signal. They allow to establish a large database of repeated and professionally read speech at low costs that is especially interesting for prosody research, but also, e.g., for concept-to-speech and socio-linguistic studies. An automatically recorded complete radio news broadcast is first segmented into individual news stories using HMM recognition. Then, the word sequence estimates of the stories are either compared directly (naive method) or realigned with the signal of other stories (realignment method) in order to find out which stories were read before and which not. Both methods can be further improved by computing "meta distances" that also take into account distances to other stories. We evaluate and compare the usefulness of the proposed methods on real life data. We find that the realignment method combined with meta distances is the most reliable of the methods and that it is well suited for the task.

1. INTRODUCTION

Today's speech and language technology research requires enormous amounts of data. As the research of the recent decade has indicated, the data should not be arbitrarily chosen but carefully selected. The process of data collection, however, is complex, time consuming and costly. We present a cost effective, automatic way of creating a database for use in research in the areas of language production, language generation, language variation as well as concept-to-speech applications.

In radio news broadcasts, especially during the night time, news stories are frequently repeated in subsequent radio news broadcasts, be it verbatim or slightly modified or reformulated. A database of such repetitions of professionally read speech is interesting for linguistic research as it can reveal syntactic, lexical, prosodic and segmental variation, as well as invariance from news story to news story. The database becomes even more interesting, when reoccuring news stories are read by multiple announcers.

Like any other linguistic model, current models of language production (cf. [1] and others) require data in order to be testable. Thanks to seminal work of Levelt's (1989), these models arrived at a degree of complexity, which becomes challenging to all areas of computational linguistics. According to Levelt, language production is processed by three general components: CONCEPTUALIZER, FORMULATOR, ARTICULATOR which use two major general knowledge sources: the LEXICON and the DISCOURSE MODEL. Figure 1 (quoted after Levelt 1989:9) gives a general flow-chart of the various processes involved in "the generation of fluent speech".

A database of reoccuring news stories allows to study the role of parts of the language generation system. Specifically, the contribution of the DISCOURSE MODEL knowledge source and the CONCEPTUALIZER processing component may be disregarded in the reoccuring news stories. This provides for a unique concentration on the variability in the function of the LEXICAL knowledge source, as well as in the particular contribution of the FORMULATOR and the ARTICULATOR processing components to the generation of fluent speech. Due to the fact that our database creation procedure selects similar news stories even if they are produced by various speakers, the phonological and phonetic variation in the process of generation of news stories can be studied systematically. Our own research concentrates on the "phonological encoding" module and particularly its part which has been dubbed "prosody generator" [1, ch. 10].

The database, in which several speakers use various formulating (encoding) and articulating (phonetic planning) strategies in order to pass on the same piece of news, is very

[*] The first author is now with Sony Intern. (Europe) GmbH, European R&D, Adv. Developments, Stuttgarter Str. 106, 70736 Fellbach, Germany

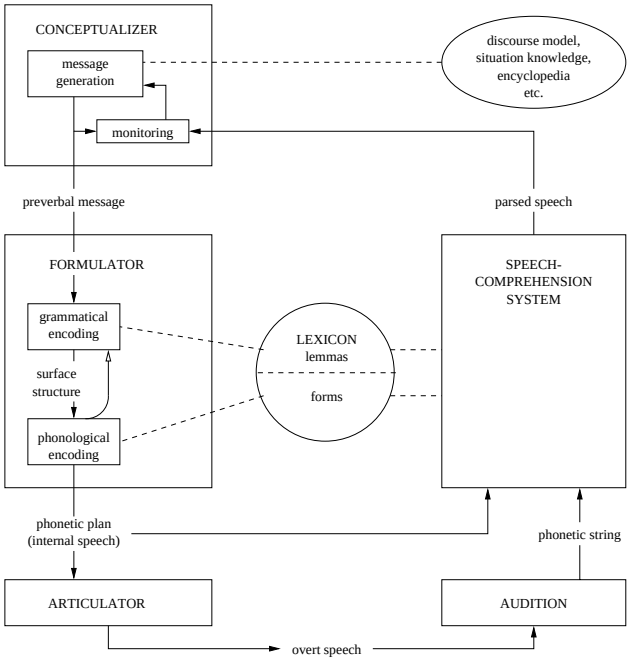

Figure 1: A blueprint of the speaker (from [1, p. 9]).

useful for prosody research. In general, a database of reoccuring news stories presented by multiple speakers is also important for concept-to-speech generation (synthesis) research as it creates a basis for studying (or automatically learning) the variation in several realizations of the same concept structure.

2. METHODS FOR FINDING REOCCURING STORIES

As a preprocessing step, the proposed methods use HMM speech recognition for segmenting a recorded complete radio news broadcast into individual news stories. The segmentation relies on the announcers' prosodic and lexical marking of boundaries between news stories and the rather fixed layout of a radio news broadcast. A simplified version of the task grammar used for segmentation into news stories is depicted in fig. 2. For a separation of the weather forecasts, the lexical markers (spotted keywords) as shown in fig 2 are used, while for a separation of the individual news stories (i.e. about five or ten per show), only prosodic marking (speech pauses exceeding a certain threshold) is exploited.

We have developed two methods, the naive method and the realignment method, for finding same or almost same news stories in the hourly radio news broadcasts. Both methods detect reoccuring news stories in subsequent news broadcasts spoken by the same or different announcers. Figure 3 shows which representations of the news stories are compared by the two methods.

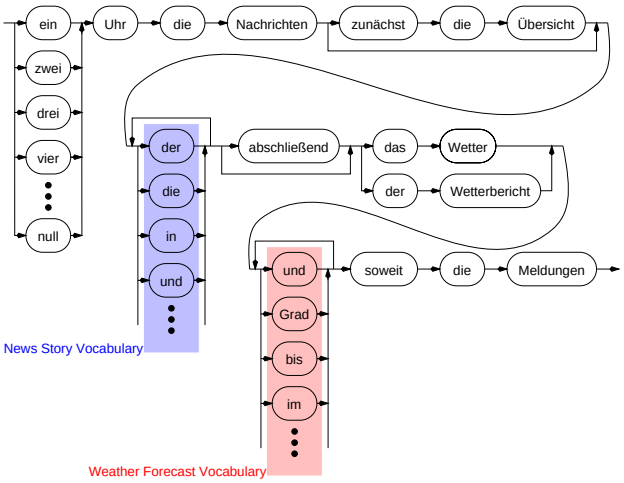

Figure 2: Simplified version of the finite state grammar for segmentation of a news broadcast into news stories.

2.1. Naive method

As a by-product of the segmentation of a complete news broadcast into individual news stories we get a word sequence estimate for every story. If the recognition is sufficiently dependable, we can assume that the sequence estimates are quite similar for reoccuring stories while for different stories they are more or less different. Hence, it is quite natural to compare the edit distance between word sequence estimates in order to find out whether the news stories from which they were estimated were (almost) identical or different. The measure applied here is the same as the well known "accuracy" of speech recognition evaluation, the only difference being that instead of comparing one recognizer output to a (correct) reference, two recognizer outputs are compared to each other. As, in general, the two word sequence estimates have a different number of words, we get two accuracy values for every pair of news stories depending on which of the two sequences is taken as the reference. By calculating the two 'accuracy' values for every pair of news stories we can simply judge on the similarity of two news stories by comparing both to a fixed threshold. This corresponds to setting every entry of the distance matrix that lies above it to 1, every other entry to 0. This new binary matrix defines a relation $R \in (news \times news)$ and, if the similarity measure and the

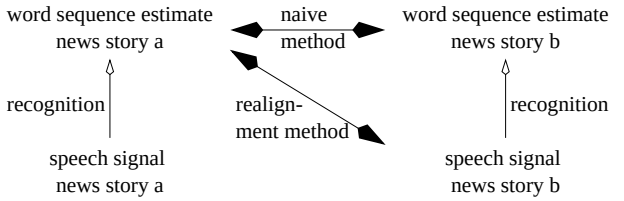

Figure 3: The both methods to compare two news stories.

threshold are chosen well, we can hope that R is an equivalence relation that partitions the news stories into clusters. In practice, we will have to face a situation in which a news story a from cluster A and a news story b from cluster B are similar enough to be above the threshold, hence $(a, b) \in R$ although a and b are different. Let us assume that all other news story pairs in A are in R, and all the pairs of B-news stories are in R as well. Then we have two options. The first is to collapse the clusters A and B since they are connected through a and b. This would correspond to making R an equivalence relation by computing the transitive hull of R. The other option is to consider aRb as a misrecognition and keep A and B as different clusters. This second option corresponds to making R an equivalence relation by finding cliques or biconnected components in the graph defined by R.

2.2. Realignment method

The naive method does not take into account the fact that some words are more easily confused than others. One typical recognition error is a substitution of a word by a *similar sounding* one. Another is the misrecognition of an OOV-word by breaking it apart into several in-vocabulary-words. These problems are particularly to be expected in our setting due to the open vocabulary problem that recognition of radio news stories generally has to face. One way to compensate these misrecognitions in the naive approach would be to find a measure that describes the similarity between words and weigh the errors accordingly. However, the development of such a measure would be a bit of an effort. Hence, instead of scoring the similarity between the words of news story a and the words of news story b on an abstract symbolic level, we estimate their similarity by considering the similarity of the words of news story a with the actual speech signal of news story b through a forced alignment. Forced alignment maps a given sequence of HMMs onto an observation sequence by using (1) information about the sequence of speech sounds of news story a as contained in, or rendered by, the word sequence estimate found during segmentation, (2) phonetic knowledge contained in the HMMs, and (3) news story b's parameterization of the speech signal. There is also a probabilistic view to the method: During the word sequence estimation for news story a, that sequence of words was found that has (according to the used modeling) the highest probability of emitting news story a's speech signal representation. By forced alignment, we calculate the probability with which the best sequence of HMMs for news story a emits the speech signal representation of news story b. The measure we used is the per frame log probability. Let us try to formalize the procedure. Let V be the (closed) vocabulary that we use for news stories segmentation, let $\mathcal{W}_i$ be sequences of words thereof ($\mathcal{W}_i \in V^*$), let $\mathcal{O}_i$ be observation sequences of news stories (the parameterization of speech).

foreach $n \in news$ **do**
  find word sequence est. $\mathcal{W}_n(\mathcal{O}_n, V)$ *by HMM-recogn.*
foreach $i \in news$ **do**
  foreach $j \in news$ **do**
    $A(i, j) = f(p(\mathcal{O}_j|\mathcal{W}_i))$

Since the costly calculation of the per frame log probability $f(p(\mathcal{O}_j|\mathcal{W}_i))$ is required for each pair of news stories, this second method is computationally more expensive than the direct symbolic approach presented in sec. 2.1.

The way to decide if two news stories are similar is, again, to define some threshold t for the per frame log probability and collect two news stories i and j in one cluster if $A(i, j) > t$ or/and $A(j, i) > t$. Again, there are two measures for every pair of news stories, for finding a partition, the graph theoretic strategies can be applied as above.

2.3. Meta distance calculation

Both methods can be improved by also taking into account the distances to other stories by calculating "meta distances": The basic distance calculations by both the naive and the realignment method for every pair of news story result in a full distance matrix (not necessarily symmetric, i.e. $A(i, j) \neq A(j, i)$ might hold) describing the similarity between news stories. Row vectors of this distance matrix describe the distance of one news story to all other stories.

The corrections that have to be made by graph theoretic algorithms indicate that it might be worth considering more information for the comparison of news story i with news story j than just looking at two elements of the distance matrix, $A(i, j)$ and $A(j, i)$. Hypothetically, if news story a_1 is very different from news story b, and a_2 is a reoccuring news story of a_1, we would expect that a_2 is quite different from b as well. And, if a_1 is not so far away from c, we would expect a_2 not to be so different from it either. Carrying over this thought to the distance matrix A, we expect that the distances of news story a_1's row in matrix A is similar to a_2's row in A, i.e., the row vectors $A(a_1, \cdot)$ and $A(a_2, \cdot)$ are somewhat nearby in $\mathbb{R}^{|news|}$. The distance between vectors in $\mathbb{R}^{|news|}$ can be measured in a variety of ways. In unsupervised clustering, an often used norm is the Euclidean distance $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$. We can use this norm (or any other) to find a new measure that describes two news stories' distance which also takes into account the distances to the other news stories. So to say, a meta distance within the "news space" can be stipulated. We can calculate such a meta distance for every pair of news stories by always taking two rows out of A so that we finally end up with a symmetric (because d is symmetric) matrix of row distances. Let us call the matrix of meta distances B, and formally define how B is calculated from A:

foreach $i \in news$ **do**
  foreach $j \in news$ **do**
    $B(i,j) = d(A(i,\cdot), A(j,\cdot))$

The calculation of B still does not free us from the requirement of defining a threshold t and it does not guarantee that R calculated from B and t is an equivalence relation, so the graph theoretic repair strategies should be applied as well.

3. EVALUATION

For evaluation and comparison of the proposed methods, three sets of German radio news broadcasts have been investigated, all were recorded via digital satellite radio from Deutschlandfunk. As the three news story sets were recorded at different days and thus do not share common stories, it is not useful to consider the sets together.

Set 1: 4 consecutive news broadcasts, manually segmented into 32 stories, manual word transliteration, checked multiple during a prosodic labelling. 15 reoccuring stories in 6 clusters.

Set 2: 14 consecutive news broadcasts, manually segmented into 110 stories, manual transliteration, multiple checked. 60 reoccuring stories organized in 23 clusters.

Set 3: 17 consecutive over night news broadcasts, automatically recorded and cut into stories. No transliterations available.

All four strategies of finding reoccuring stories (naive method,realignment method)$\times$(without/with meta distance calculation) were applied to set 1 and set 2.[1] In all evaluations, minor reformulations of news stories were tolerated, that is, news stories were counted as *same or almost same*, if the edit distance (deletions+insertions+substitutions) of the (correct) transliterations did not exceed a threshold of 10. As a correction strategy, transitive hull computation was applied. The reported results are based on an optimized threshold, although the optimal values are approximately the same for all sets (see [2] for details). The results are given in table 1. With meta distance calculation, both methods are able to perfectly separate the 6 reoccuring news story clusters of set 1; however, for the naive method, choosing the right threshold is not as easy as with the realignment method, for which scores for same news story pairs (ranging from about 80 to about 280) are very well separated from scores for different story pairs (starting from about 500). For set 2, the realignment method with meta distances misses only 4 out of 60 reoccuring stories.

Now, what could we expect as a yield from one night's news broadcasts? We used set 3 to find clusters of reoccuring stories from a fully automated collection and segmentation into stories. The naive method with meta distances gave 36 stories in 14 clusters (2x5+2x3+10x2), the realignment with meta distances 54 in 20 clusters (5x4+4x3+11x2). For the naive method, 2 clusters were wrong, after correction, (1x5+2x3+11x2) remained. In contrast, with the realignment method, all found clusters were correct except one 4 stories cluster that contains one reformulated version exceeding our ten word edit distance criterion slightly by three words.

[1]Word sequences were estimated using 3 emitting state monophone HMMs and as a vocabulary the transliterations 100 most frequent words.

	correct	naive	realign-ment	naive +meta	realignm. +meta
Set 1	3×3	$0\|2\|1$	$1\|1\|1$	$0\|3\|0$	$0\|3\|0$
	3×2	$0\|2\|1$	$1\|1\|1$	$0\|3\|0$	$0\|3\|0$
Set 2	1×8	$1\|0\|0$	$1\|0\|0$	$0\|0\|1$	$0\|1\|0$
	1×5	$0\|1\|0$	$0\|1\|0$	$0\|1\|0$	$0\|1\|0$
	5×3	$0\|4\|1$	$0\|4\|1$	$0\|4\|1$	$0\|3\|2$
	16×2	$2\|13\|1$	$2\|14\|0$	$1\|13\|2$	$0\|15\|1$

Table 1: Results for set 1 and set 2. First column: clusters according to human transliteration, $x \times y$ means x clusters of y reoccuring stories. Further columns: $a|b|c$ means b clusters correctly found, a found clusters were too large (contain an additional, non-equivalent news story), c too small (one or more of the equivalent stories was missing).

4. CONCLUSIONS

The evaluation shows that the realignment method together with the calculation of meta distances contributed in this paper is an excellent way for detecting reoccuring radio news stories. We see a potential to apply the proposed methods for further areas, but did not experimentally address them. First, it should be checked if the method could be applied to find out repetitions of OOV words in ASR. By investigating more than one occurrence of an OOV word, automatic generation of new lexicon entries might be possible with a higher transcription accuracy. Next, and coming back to radio news stories, it is possible to invert the selection procedure from the originally intended behavior. Instead of showing up repeated stories, we could suppress repetitions, so that the proposed methods could be useful for summarizing all (different) stories of the radio news broadcasts for the purpose of archiving or information retrieval applications.

References

[1] W. J. M. Levelt. *Speaking: From Intention to Articulation*. The MIT Press, Cambridge Massachusetts, 1989.

[2] S. Rapp. *Automatisierte Erstellung von Korpora für die Prosodieforschung*. PhD thesis, University of Stuttgart, Institut für maschinelle Sprachverarbeitung, 1998.