Heuristic Syllabification and Statistical Syllable-Based Modeling for Speech-Input Topic Identification *

Pierre Ouellet[1] and Pierre Dumouchel[1,2]

[1] Centre de recherche informatique de Montréal (CRIM)
Pierre.Ouellet@crim.ca
[2] Ecole de technologie supérieure
Pierre.Dumouchel@crim.ca

Abstract. We describe a heuristic syllabification method and the use of a statistical syllable n-gram language model for discriminating between a closed set of topics. The syllabification method works by assigning costs to consonant clusters and then splitting the clusters where the cost is minimized. We apply the syllabification on a pronunciation dictionary which maps words to phone sequences; the result is then used for producing an inventory of syllables and to transform word- or phone-level transcriptions into syllable-level transcriptions. In topic identification (TID), each topic is represented by a backoff syllable n-gram statistical model trained on a sample of syllabified texts deemed representative of the topic. Identifying the topic of an unknown utterance is performed by transcribing it into syllables using a speech recognizer; the transcription is then evaluated (scored) with respect to each topic model, and the topic for which the score is highest is deemed the most likely. The training texts for topic models can be produced manually, or automatically using speech recognition. We evaluate the accuracy of the syllabification by referring to the Celex lexicon, which includes syllabifications; topic identification performance is evaluated on a subset of the Switchboard corpus.

1 Introduction

Searches on the World Wide Web are a familiar form of document retrieval: a number of keywords are entered into a search engine, and pages containing the keywords are returned to the user for browsing. This method is limited to textual information, which is only part of the information found on the Web. Images, audio and video can contain useful information, and unless these are explicitly indexed or annotated with text, Web searches cannot find the information in these types of data. Hence, much research concentrates on retrieving unindexed and unannotated audio or video documents using non-textual criteria. This paper concentrates on topic identification from speech input, that is, determining

* This work was funded by the Department of National Defence of Canada

the general topic of a series of spoken utterances by processing automatically the audio data itself using Automatic Speech Recognition (ASR) and evaluating the resulting transcription with respect to different topic models, to find the most likely topic underlying the given speech data.

2 Preliminaries

2.1 Automatic Speech Recognition

An ASR system consists of a mathematically describable model of speech, that is, an estimated correspondence between spoken language and written forms which can be evaluated by computer software. Given a spoken utterance in digital form, an ASR system produces a (possibly imperfect) transcription of it. Such a system can be decomposed into two parts:

1. An *acoustic model*, which describes a correspondence between basic speech (concrete) units and basic linguistic (abstract) units.
2. A *language model*, which describes relationships between basic linguistic (abstract) units.

The linguistic units are typically words, syllables, or phones; the application dictates the most useful choice. The speech units are usually context-dependent phones, that is, each abstract phone can be modeled by multiple allophone models, the choice of which to use depending on the surrounding phones. If the abstract and concrete units are not at the same level, additional information is needed to establish a correspondence, for example a pronunciation lexicon which maps words or syllables to phone sequences. The ASR system we use here uses context-dependent phones as speech units and syllables as linguistic units. The lexicon maps each syllable to its phone sequence.

2.2 Language Modeling

The language model of a typical modern ASR systems is quite simple: a probabilistic prediction is made about which linguistic unit is to follow, given the $n-1$ previous linguistic units (n-gram LM). The advantages of such a simple LM are that it is mathematically sound, and that it is tractable (due to the locality of the information). For example, word bigrams and trigrams ($n = 2$ and $n = 3$) are typically used in large vocabulary ASR applications such as dictation systems. The probabilities are estimated from large text corpora, which should be chosen to be as similar as possible in style as the language of the target application. Dictation systems often use LM's trained on years of newspaper articles. In our TID system, the ASR language model is a syllable trigram and is estimated from telephone conversation transcriptions provided with the Switchboard corpus (described later).

2.3 Topic Modeling

We assume that if we train an LM (estimate its probabilities) using texts pertaining to a specific topic, the specific probabilities will somehow be biased towards the topic. Given multiple topic-dependent LM's and using conditional probabilities, we can evaluate the *a posteriori* probability of each topic given a new text, and thus determine the most likely topic for said text. We can use topic LM's trained in two ways:

- *Supervised*, that is, using manual transcriptions of topic-specific speech.
- *Unsupervised*, that is, using ASR transcriptions of topic-specific speech.

Supervised training usually yields more accurate results, but unsupervised training is more practical since it does not require manual transcriptions.

2.4 The Switchboard Corpus

In the early 1990's, a corpus of telephone conversations was collected by Texas Instruments with funding from DARPA. Callers dialed an automated system which would match them with another caller after having suggested a topic for discussion. The result is a corpus of spontaneous yet polite speech, where conversations usually last between five and ten minutes (people were cut off after a certain limit). It is notoriously difficult to use ASR techniques successfully on Switchboard, because of the multiple sources of variability: multiple types of telephone handset microphones, local and long-distance calls, noise, different accents, etc.

2.5 ASR Errors and Accuracy

An ASR system can make three types of errors (we assume here that the linguistic unit is the word):

1. *deletion*: an uttered word is not present in the ASR transcription.
2. *insertion*: a word not uttered is present in the ASR transcription.
3. *substitution*: a word is uttered and another is present in its corresponding place in the ASR transcription.

There can be many combinations of errors which lead from a reference transcription to the ASR transcription; for example, a deletion followed by an insertion can produce the same result as a substitution. We decide on one specific combination by assigning a cost to each type of error, and choosing the combination of errors which leads from the reference transcription to the ASR transcription with minimum cost. We define accuracy as $Acc = \frac{C-I}{N}$, where N is the number of linguisitic units in the correct transcription, C is the number of units correctly recognized, and I is the number of insertions. Our syllable recognizer produces transcriptions which are overall about 43% accurate at the phone level.

3 System Description

Given our ASR system and a set of n-gram topic models, determine the topic of a given spoken utterance thus:

1. Produce a syllabic transcription of the utterance using the ASR system.
2. Evaluate the likelihood of each topic model with respect to the transcription produced, and choose the most likely topic as the correct one.

4 Introducing Syllables

4.1 Motivation

We wish to improve the accuracy of topic modeling by using syllables instead of phones. Introducing syllables into the system is relatively easy; the harder part is determining which set of syllables to use. We chose to syllabify a pronunciation dictionary well known in ASR circles, `CMUDict-0.6`. The word-level transcriptions from Switchboard were then replaced by the corresponding syllable-level transcriptions using the syllabified lexicon, and all syllables occurring in the syllabified transcriptions were included in the syllable lexicon, for a total of 7,387 syllables.

4.2 Syllabification Method

We define a syllable as a possibly empty consonant cluster (CC), followed by a vowel, followed by a possibly empty CC; this simple but practical definition will suit our needs. We then reduce the problem of syllabification to the problem of choosing where to split CC's between any two consecutive vowels. We distinguish among initial CC's (onset) and final CC's (coda), each of which can be "good" or "bad". We define that a CC is bad until proven good. The proof is obtained by examining a pronunciation dictionary mapping words to phone sequences (the actual words are not relevant here): CC's found frequently at the beginning of words are deemed good initial CC's; similarly, CC's found frequently at the end of words are deemed good final CC's. The exact threshold for "frequently" is determined empirically. We then define a function which assigns a cost to each type of CC. One important and somewhat arbitrary criterion is that long onsets are preferred to long codas. Also, we chose to limit good CC's to at most three consonants. The cost function is ak, where k is the number of consonants in the cluster, and a has a value according the the following table:

CC type	value of a
good initial	0
good final	1
bad initial	4
bad final	5

Good CC's (initial or final) of length three or less are preserved (not split), since they cost at most 3 units, whereas a bad CC costs at least 4 units. For any given sequence of consonants between two vowels, all possible splits into CC's are performed. The cost of the split under consideration is the cost of the final CC belonging to the previous syllable plus the cost of the initial CC belonging to the following syllable. The split of lowest cost is then chosen over all possible splits.

4.3 Evaluation

We have tested our syllabification method on a pronunciation lexicon which includes syllabifications, CELEX [Baayen93]. The process used for creating these is unknown, but probably involves applying a set of linguistic rules and veryfing the results manually. We chose the (British) English word-forms lexicon, which contains 148,329 unique word/pronunciation entries, some of which are phrases. For 83.1% of the entries, the syllabification produced was the same as that provided with the lexicon. Monosyllabic entries (6.2%) were of course always syllabified correctly (i.e., not at all). When we consider the individual CC splits, our method split CC's in the same place as the reference syllabifications for 91.9% of the syllable breaks. The percentages reported do not take into account the frequency of occurrence of words.

5 Topic Identification

5.1 Experimental Setup

The test set consists of 507 speech files (conversation sides) from Switchboard, each related to one of ten topics, for a total of about 25 hours of speech. These files were split into 10 test sets (10% each); to each of these, a topic training set consisting of the remaining 90% of the files were associated. The result is 10 test configurations, each of which comprises 50 or 51 test files and 456 or 457 training files. The ASR system is a syllable recognizer whose output is used for both phone- and syllable-based topic models, decomposing syllables into their constituent phones where necessary.

5.2 Results

The best phone accuracy for the syllable-based ASR system is 43.6%.

Supervised vs. Unsupervised Topic Training Using known transcriptions for topic model training, we were able to identify the topic correctly for 63.7% of the test files. Using instead ASR-produced transcriptions, the correct topic was selected 57.6% of the time. Therefore, the errors in the ASR transcription have a significant impact on topic identification rate (6.1% absolute difference). Moreover, if the ASR system had 100% accuracy, we could achieve 92.1% TID rate on this test set. This indicates that ASR accuracy is even more crucial when transcribing test files.

Syllable Topic Models vs. Phone Topic Models If we use a phone 4-gram unsupervised topic model instead of an unsupervised syllable bigram topic model, we can obtain 57.4% TID rate (-0.2%). This indicates that the advantage of using syllables for topic modeling is slim, provided that the span of the n-grams is roughly the same number of phones on average, and that the number of statistical parameters is comparable.

6 Conclusion

We have described the introduction of syllables into our phone recognizer, and as linguistic units for topic modeling. In particular, we showed how a simple heuristic syllabification method could syllabify a given lexicon with about 90% accuracy. A slight improvement has been observed when using syllables for TID, although it may not be significant. We expect that a more detailed study might yield a greater increase in accuracy.

7 Acknowledgements

The work on topic identification (including syllabification) was funded by the Department of National Defence of Canada. We wish to thank Paul Soble and Karl Boutin for their support.

References

[Baayen93] Baayen, R. H., Piepenbrock, R., van Rijn, H. "The CELEX Lexical Database (CD-ROM)". Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1993.