INTEGRATION OF SEVERAL INFORMATION SOURCES FOR ROBUST CLASS-BASED STATISTICAL LANGUAGE MODELLING

Géraldine Damnati
France Télécom, CNET DIH/DIPS, 22307 Lannion Cedex, France.
geraldine.damnati@cnet.francetelecom.fr

ABSTRACT
The lack of training material, which is particularly problematic in the context of human-machine spoken dialogue, is addressed in this paper. Automatic classification is known to be useful to model spontaneous speech, but contextual methods are, by nature, sensitive to the data. In test conditions similar to training, the heterogeneity of classes does not matter, as long as they faithfully reflect training. When considering some robustness issues, such as domain or vocabulary adaptation, consistency of classes becomes relevant, and classical contextual algorithms can not provide this when faced to a low amount of training data. Hence, this paper proposes a new theoretical framework for hierarchical clustering, where several information sources can be combined at the similarity criterion evaluation level. This allows a priori information to be taken into account as a complement of the contextual information. Consistent classes can be built, while remaining suited for spontaneous speech modelling.

1. INTRODUCTION
Class-based language models are now commonly accepted as a helpful approximation, with respect to the sparse data issue. Collecting training data is all the more problematic in the context of human-machine spoken dialogue, since the transcription of real utterances is an expensive task. Thus, the language model of the speech recognition system of a dialogue application, has to be built from a low amount of training data, and classes are particularly useful in this context.

As studied before [1,2], spontaneous speech is very peculiar and implies specific treatments. The poor syntactic structure of speaker utterances, for example, discredits the use of fully a priori classification techniques, such as POS tagging. Automatic classification techniques, relying on contextual information extracted from training data, are thus more appropriate to take into account the particularities of spontaneous speech.

In order to automatically build classes, several clustering algorithms have been proposed and compared so far [3]. Through what follows, emphasis is placed on bottom-up hierarchical clustering, leading to the definition of a new clustering criterion, which combines several information sources.

In section 2, the hierarchical clustering issue is redefined from the data analysis point of view. In this framework, the classical hierarchical clustering algorithm [4] used for language modelling, happens to be a particular case corresponding to simultaneous classification. Therefore, a more general framework is introduced in section 3, which relies on the distinction between predecessors and successors. After presenting the limitations of the simple contextual information, section 4 proposes an extension, which allows several information sources to be combined within the clustering process. Such a richer criterion leads to more consistent and robust classes.

2. HIERARCHICAL CLUSTERING
2.1 General definitions
In this work, classification is apprehended from its fundamentals, that is to say from the data analysis point of view [5]. The baseline of any analysis is the description of observations through a data table. Classification thus consists in simplifying the representation of data, with the minimum loss in precision for the description of these data. Suppose one observes a set of individuals, each of them being described by a set of characteristics. Representing two of these characteristics by random variables X and Y, it is possible to summarise the relationship between X and Y, by crossing their modalities through a contingency table.

Such a table contains at line i and column j, the number of individuals observed with simultaneously both characteristics x_i and y_j. Considering the example of individuals characterised by the colour of their hair (X) and the colour of their eyes (Y), if x_i corresponds to the modality "dark" and y_j corresponds to the modality "blue", the element in the table, at line i and column j, would be the number of individuals having dark hair and blue eyes.

When dealing with bigram language modelling, individuals are pairs of words, observed in a training corpus. Building the data table consists in counting these pairs, and in reporting at line i and column j, the number of pairs whose left part is the word w_i, and whose right part is the word w_j. Both rows and columns denote the words of the vocabulary, thus both X and Y modalities are represented by the same symbols, which is source of confusions. It is somehow relevant to distinguish between rows and columns, by calling X the predecessor variable and Y the successor variable.

Given such a representation of the data, if the analysis holds for Y variable, classification consists in finding similarities between Y modalities, according to their co-occurrence with X modalities. In particular, hierarchical clustering consists in building a binary tree from the set of modalities of a variable Y, according to their
similarity with respect to the variable X. When building classes for a bigram LM, the observations are the possible pairs of words, while X and Y respectively stand for predecessor and successor words in these pairs. A merging criterion, based on information theory, is defined as the minimum loss of average mutual information I(X;Y) between variables X and Y.

2.2 Simultaneous clustering

The application of hierarchical clustering for language modelling purpose, was introduced in [4]. In fact, the proposed algorithm happens to be a particular case in the hierarchical clustering theoretical framework, where both rows and columns of the data matrix are clustered. As a matter of fact, the classification algorithm relies on the way classes are used in the classical bi-class approximation for probability computation:

\[ P(y|x)\varphi P(y|\mathcal{C}(y))P(\mathcal{C}(y)|\mathcal{C}(x)) \]  

According to this formulation, no distinction is made between \( x \) and \( y \) from the class set point of view. In that case, the problem is to build one mapping of the words \( (w\in\mathcal{C}(w)) \), with no respect to the fact that \( w \) can be a predecessor or a successor word. As a consequence the clustering algorithm has to hold both for rows and for columns of the data table, that is both for X and for Y modalities.

3. SEPARATE CLUSTERING

In order to differentiate X and Y variables, this paper proposes to redefine the bi-class approximation by making the distinction between predecessors and successors:

\[ P(y|x)\varphi P(y|\mathcal{C}(y))P(\mathcal{C}(y)|\mathcal{C}(x)) \]  

According to this formulation, the problem is then to build two different mappings of the vocabulary, one for words as predecessors \( (w\in\mathcal{C}(w)) \) and the other for words as successors \( (w\in\mathcal{C}(w)) \).

Other approaches make the distinction between predecessors and successors, in terms of state mapping and category mapping [6], or in terms of asymmetric model for a priori classification [7]. As for the present work, the distinction between the two aspects of classification results naturally from the hierarchical clustering framework.

The theoretical justification agrees anyhow with more practical considerations. According to simultaneous classification, words are grouped if they share similar left and right context. According to separate classification, two successors are grouped if they occur with similar left contexts while two predecessors are grouped if they occur with similar right contexts. In fact, as pointed out in [7], it seems clear enough that some words can appear with similar left context, and that grouping them can therefore be helpful in order to predict them. On the other hand, the same set of words can appear with different right context, hence grouping them can happen to be harmful for the prediction of the next word.

In this paper, the separate classification algorithm is described for Y modalities clustering. That is, only successors are grouped, leading to the following model, corresponding to the word-class interaction studied in [8] in another context:

\[ P(y|x)\varphi P(y|C(y))P(C(y)|x) \]  

The classification extends, by symmetry, to the clustering of X modalities, in which case the approximation would be: \( P(y|x)\varphi P(y|C(x)) \).

3.1. Distance and complexity

The distance between two modalities \( y_j \) and \( y_k \) of Y, is the loss of average mutual information \( I(X;Y) \) induced by the merging of these modalities.

Let \( \varphi(x,y) \) denote:

\[ \varphi(x,y) = P(x,y) \log \frac{P(y|x)}{P(y)} \]

The average mutual information between variables X and Y is then defined as:

\[ I(X;Y) = \sum_{x,y} \varphi(x,y) \]  

Suppose two modalities \( y_j \) and \( y_k \) are clustered into a new modality \( y_r \), leading to a new random variable \( Y_r \), then the distance between \( y_j \) and \( y_k \) is the difference between \( I(X;Y) \) and \( I(X;Y_r) \):

\[ \Delta(y_j,y_k) = \sum_x \varphi(x,y_j) + \sum_x \varphi(x,y_k) - \sum_x \varphi(x,y_r) \]  

The merging criterion is therefore the minimum loss in average mutual information.

When considering simultaneous clustering of X and Y modalities, the computational cost of the algorithm is essentially due to the distance matrix updating. As a matter of fact, merging two classes \( c_1 \) and \( c_2 \) affects the distances between every pairs of words. In other words, a complete distance matrix updating has to be achieved at each iteration. This is an important drawback when the size of the vocabulary grows. On the contrary, when dealing with successor clustering, the matrix updating only concerns the distance between words and the new class \( y_r \). This implies a significant reduction in computational cost.

3.2. Limitations of the contextual criterion

Even though automatic classification is more appropriate to model spontaneous speech, its construction remains problematic as it is very sensitive to the lack of training material. What’s more, peculiarities of spontaneous training corpora are seriously harmful for data-based clustering algorithms. Among such peculiarities, one can observe the high proportion of short utterances, leading to a high representation of the \(<\text{start/endpoint}\>\) context. Furthermore, the low amount of available spontaneous data has several consequences. Some words are unseen in the training set and are thus excluded from the clustering process. Many low occurring words are observed with a very low variety of contexts, and can thus be badly clustered.
This results in a bad repartition of the words among the different classes. Some classes gather a large amount of low occurring words, corresponding to common left contexts, such as "the" or "<start>". As a consequence, a class can contain words of different syntactic or semantic nature. The classes content is not a relevant argument when the objective is the perplexity reduction in conditions similar to training. Yet, if one is concerned with more general problems such as portability of the classes towards a new domain, or introduction of new words into the classification, the classes consistency becomes a relevant characteristic. In such conditions, limitations of the simple contextual criterion, due to spontaneous speech peculiarities, are particularly problematic.

In order to overcome the lack of training data, this paper proposes a way to extract several information sources from the training corpus and, furthermore, to combine them into a new, enriched, hierarchical clustering theoretical framework. Instead of increasing the amount of data, this method consists in increasing the amount of information extracted from the data.

4. TOWARDS ROBUST CLASSIFICATION: COMBINING INFORMATION SOURCES

4.1. Extracting various information sources

Considering separately X and Y modalities leads to several remarks, summarised in the following question:

"Would it be possible to characterise Y modalities according to richer information than the simple observation of the preceding word?"

As a matter of fact, the preceding word is not the only information source that can be extracted from the data. This method consists in increasing the amount of information provided by the joint observation of the two preceding words. Similarity between successors is then related to their occurrences in a trigram context. This results in a better distribution of words into classes, as the variety of possible contexts is enriched. Now, if the objective is to improve the consistency of classes from a syntactic or semantic point of view, it is interesting to take into account a priori information. To do so, the training corpus is tagged beforehand, and X is the tag of the successor observed in the training corpus.

The notion of individual defined in section 2, is now modified. If the additional variable is the tag \(\tau\) of the current word \(m_i\), an observed individual is an occurrence of \((X_1, X_2, Y)\) and can be denoted \((m_{i-1}, \tau, m_i)\).

The joint mutual information is then:

\[
I(X_1, X_2; Y) = \sum_{x_1, x_2, y} \log \frac{P(x_1, x_2, y) P(y) \prod_{i=1}^n \frac{P(x_{i-1}) P(x_i | x_{i-1}, y) P(y | x_{i-1}, y)}}{P(y)}
\]

This value quantifies the information on successors provided by the joint observation of \(n\) sources.

4.3. Combining two information sources

The first experiments have been carried out with two information sources.

Let \(\phi(x_1, x_2, y)\) denote:

\[
\phi(x_1, x_2, y) = P(x_1, x_2, y) \log \frac{P(y | x_1, x_2)}{P(y)}
\]

The distance between two modalities \(y_j\) and \(y_k\) of Y, merged into \(y_s\), is therefore:

\[
\Delta(y_j, y_k) = \sum_{x_1, x_2} \phi(x_1, x_2, y_j) + \sum_{x_1, x_2} \phi(x_1, x_2, y_k) - \sum_{x_1, x_2} \phi(x_1, x_2, y_s)
\]

In order to remain suited to spontaneous speech bigram language modelling, variable \(X_1\) still corresponds to the preceding word. As for variable \(X_2\), two choices are described in this paper, corresponding to two different objectives.

If the objective is to improve the LM perplexity on the test set (which represents conditions closed to training conditions), it seems natural to try to extract larger contextual information from the data. Therefore, the variable \(X_2\) corresponds to the observation of the next to last word. Similarity between successors is then related to their occurrences in a trigram context. This results in a better distribution of words into classes, as the variety of possible contexts is enriched. Now, if the objective is to improve the consistency of classes from a syntactic or semantic point of view, it is interesting to take into account \(a\ priori\) information. To do so, the training corpus is tagged beforehand, and \(X_2\) corresponds to the tag of the successor observed in the training corpus.

The notion of individual defined in section 2, is now modified. If the additional variable is the tag \(\tau\) of the current word \(m_i\), an observed individual is an occurrence of \((X_1, X_2, Y)\) and can be denoted \((m_{i-1}, \tau, m_i)\).

The joint mutual information is then:

\[
I(X_1, X_2; Y) = \sum_{x_1, x_2, y} \log \frac{P(m_{i-1}, \tau, m_i | x_1, x_2, y) P(m_{i-1}, \tau, m_i | x_1, x_2, y) P(m_{i-1}, \tau, m_i | x_1, x_2, y)}{P(m_{i-1}, \tau, m_i | x_1, x_2, y)}
\]

With this additional variable, successors are grouped if they share both similar left contexts and the same tag. It amounts to a constraint on possible merges, as in [10], but the constraint is included in the criterion, and the disambiguation problem is avoided as the tag is observed in the context. This allows more consistent classes to be built.

5. EXPERIMENTS

Experiments are run on a spontaneous speech corpus composed of utterances of human-computer spoken dialogues, on the AGS voice service directory inquiry demonstrator [11]. The vocabulary contains 878 different words, 756 of them occurring in the training text. The training corpus consists of 39673 occurrences of words (7863 sentences), and the test corpus consists of 3584 words from 724 sentences.

Speech recognition tests are carried out with a HMM-based, speaker independent, continuous speech recognition software working over the telephone network. The acoustic model is mono-gaussian.
In order to extract a priori information from the data, the training corpus has been tagged, using two different taggers. One is a fully rule-based tagger [12]. It is an extension of a phonetic transcriber, first devoted to text-to-speech synthesis and, later on, to speech recognition. The second corresponds to a mixed approach where rules have been introduced into a probabilistic tagger [13]. Both lead to equivalent performance in term of the resulting bi-class LM perplexity. The following results correspond to the use of the mixed tagger.

Four models are compared in the following table and figure, in terms of test set perplexity and word error rate. A word bigram LM, smoothed by "absolute discounting", a bi-class LM with 400 classes build by simultaneous clustering. The two others are derived from equation (3), the classification of successors being obtained through the combination of two variables. The first additional variable is the next to last word, and the other is the tag of the current word. Class-based models are smoothed by "absolute discounting" and the number of classes, optimised on a separate validation set, is given in Table 1.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Test Set Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>word bigram</td>
<td>22.54</td>
</tr>
<tr>
<td>simultaneous classification (400)</td>
<td>22.56</td>
</tr>
<tr>
<td>successors: prec. word + current tag (450)</td>
<td>22.46</td>
</tr>
<tr>
<td>successors: 2 prec. words (350)</td>
<td>22.31</td>
</tr>
</tbody>
</table>

In Fig. 1, speech recognition word error rates are given as a function of the active search space percentage during decoding. A low active search space corresponds to strict pruning conditions, while a high active search space corresponds to less restrictive pruning conditions. In the former conditions only a small subset of very likely solutions are explored. Thus, observing the various models behaviour, indicates how well the likely events fit with the good solutions. Strict conditions are important to take into account, when one needs to provide a word-graph as the output of the speech recognition system. In fact, the search space size conditions the word-graph size, which is a crucial issue.

6. CONCLUSION

This paper proposes a new approach in hierarchical clustering for statistical language modelling. Separate clustering has been introduced as an alternative to the classical simultaneous clustering. Furthermore, the distinction between predecessors and successors leads to the definition of a new theoretical framework, allowing to combine several information sources, at the clustering process level. This method circumvents the lack of training data issue, by attempting to extract more information from the available data. As a priori knowledge can be introduced into automatic classification, the obtained classes are more consistent syntactically and semantically, while keeping track of the contextual information, and thus remaining relevant for spontaneous speech modelling. Such classes provide an efficient baseline for the vocabulary adaptation issue, when a dialogue task is to evolve, as they are more suited for the introduction of a new word.

BIBLIOGRAPHY