A semi-automatic structure learning method for language modeling

Vitor Pera

Faculdade de Engenharia da Universidade do Porto
Rua Dr Roberto Frias, s/n 4200-465 Porto, Portugal
http://www.fe.up.pt
vcp@fe.up.pt

Abstract. This paper presents a semi-automatic method for statistical language modeling. The method addresses the structure learning problem of the linguistic classes prediction model (LCPM) in class-dependent N-grams supporting multiple linguistic classes per word. The structure of the LCPM is designed, within the Factorial Language Model framework, combining a knowledge-based approach with a data-driven technique. First, simple linguistic knowledge is used to define a set with linguistic features appropriate to the application, and to sketch the LCPM main structure. Next an automatic algorithm selects, based on Information Theory solid concepts, the relevant factors associated to the selected features and establishes the LCPM definitive structure. This approach is based on the so called Buried Markov Models[1]. Although only preliminary results were obtained, they afford great confidence on the method's ability to learn from the data, LCPM structures that represent accurately the application's real dependencies and also favor the training robustness.

Keywords: language modeling, structure learning, class-dependent N-gram

1 Introduction

The ability of the language model (LM) to represent with enough accuracy the real linguistic structure and redundancy patterns present in an application, reducing properly and as much as possible the task perplexity, is in general crucial for the performance of the system using the LM. N-grams continue to be quite common, at least in automatic speech recognition (ASR), given their effectiveness in many applications and also because linguistic expertise is dispensed[8]. Nevertheless, when the vocabulary is very large the sparse data estimation problem usually becomes critical. Statistical modeling techniques based on data sharing or smoothing principles, e.g. back-off strategies or interpolation methods, have been developed to mitigate over-fitting effects[3, 5]. Another proposed approach has been the class-dependent N-grams. These are at the basis of this work. It has been recognized that in the case of some applications exhibiting relatively complex linguistic patterns involving multiple linguistic features, new and better approaches to exploit those patterns are needed[6, 7]. This work addresses this particular issue, proposing a method to optimize according to some criteria, based on solid Information Theory principles, the structure of the linguistic classes prediction model.

The structure of the paper is as follows. Section 2 presents a brief analysis of the class-dependent N-grams modeling ability. Section 3 begins with some discussion on the application's properties motivating the proposed method, and then presents its two main steps. Preliminary results obtained using this method are presented in section 4. The main conclusions of this work are pointed-out in section 5.

2 The class-dependent N-grams

Given a sequence of words $\omega_{1:T}$, the Language Model estimates the probability $P(\omega_{1:T})$, which can be factorized as $\prod_t P(\omega_t|\omega_{1:t-1})$. Let assume that each word, ω, in the vocabulary, $\mathcal{V}$, is associated to some subset, $C(\omega)$, eventually a singleton, of the linguistic classes set, $\mathcal{C}$. Then

$$P(\omega_t|\omega_{1:t-1}) = \sum_{c_t \in C(\omega_t)} P(\omega_t|c_t, \omega_{1:t-1}) P(c_t|\omega_{1:t-1}). \tag{1}$$

Two assumptions are made that approximate this conditional probability: 1) it depends almost entirely of the recent history, so N and M values are set for ω and c depths, respectively; 2) the linguistic classes prediction can be adequately modeled discarding the word terminals information. Accordingly,

$$P(\omega_t|\omega_{1:t-1}) \approx \sum_{c_t \in C(\omega_t)} P(\omega_t|c_t, \omega_{t-N+1:t-1}) P(c_t|c_{t-M+1:t-1}). \tag{2}$$

In general $M > N$, typical ranges are $N = 2, ..., 5$ and $M = 3, ..., 7$ (and $|\mathcal{C}| << |\mathcal{V}|$).

Let now make a brief analysis of the class-dependent N-grams modeling ability, comparing it with standard N-grams. Let consider $|C(\omega)| = 1, \ \forall \omega \in \mathcal{V}$, which is accurate for many words, in order to simplify the following analysis. Accordingly,

$$P(\omega_t|\omega_{1:t-1}) \approx P(\omega_t|c_t, \omega_{t-N+1:t-1}) P(c_t|c_{t-M+1:t-1}), \tag{3}$$

where $c_t = f(\omega_t)$ for some known function f. Then, the conditional probability expected log-value $\mathcal{E} = E[log(P(\omega_t|\omega_{1:t-1}))]$, over a representative data set $\{\omega_{1:T_l}\}_{l=1}^{L}$, can be approximated as follows (assuming $M > N$):

$$\begin{aligned}
&\sum_{\omega_{t-M+1:t}} P(\omega_{t-M+1:t}) log(P(\omega_t|c_t, \omega_{t-N+1:t-1}) P(c_t|c_{t-M+1:t-1})) \\
&= \sum_{\omega_{t-M+1:t}} P(\omega_{t-M+1:t}) log\Big(\frac{P(\omega_t, c_t|\omega_{t-N+1:t-1})}{P(c_t|\omega_{t-N+1:t-1})} P(c_t|c_{t-M+1:t-1})\Big) \\
&= \sum_{\omega_{t-N+1:t}} P(\omega_{t-N+1:t}) log P(\omega_t|\omega_{t-N+1:t-1}) \\
&+ \sum_{\omega_{t-M+1:t}} P(\omega_{t-M+1:t}) log \frac{P(c_t|c_{t-M+1:t-1})}{P(c_t|\omega_{t-N+1:t-1})}
\end{aligned}$$

The first term expresses the conditional probability expected log-value corresponding to a standard N-gram. Continuing to assume that c is uniquely determined by ω, the second term can be written:

$$\sum_{\omega_{t-M+1:t}} P(\omega_{t-M+1:t}) log \frac{P(c_t|c_{t-M+1:t-N}, c_{t-N+1:t-1})}{P(c_t|c_{t-N+1:t-1})}$$
$$= \sum_{\omega_{t-M+1:t}} P(\omega_{t-M+1:t}) log \frac{P(c_t, c_{t-M+1:t-N}|c_{t-N+1:t-1})}{P(c_t|c_{t-N+1:t-1})P(c_{t-M+1:t-N}|c_{t-N+1:t1})}$$

It is clear that the second term represents the conditional mutual information between c_t and $c_{t-M+1:t-N}$ given $c_{t-N+1:t-1}$, i.e., $I(c_t; c_{t-M+1:t-N}|c_{t-N+1:t-1})$. Comparing with a standard N-gram, is fair to expect that the descriptive power of this model is substantially larger, iff $c_{t-M+1:t-N}$ conveys relevant information about the outcome of c_t, not present in $c_{t-N+1:t-1}$. Some quite common circumstances favor this potential improvement: 1) the value of N is in general severely limited by the data resources, therefore not allowing to capture important past cues; 2) the contrary occurs in relation to M, which value in general can be made large enough to model early cues that can be useful; 3) in many real applications the entropy associated to the conditionals $P(c_t|c_{1:t-1})$ is small, which favors the linguistic classes information as an aid to predict the sentence words.

3 The linguistic classes prediction model

3.1 A factorial language model approach

The linguistic classes prediction model (LCPM) design follows the factorial language model (FLM) formalism. In terms of notation, the linguistic classes variable c becomes a vector with K components (factors), $f^{1:K}$; accordingly, hereafter, $c_{t_1:t_2}$ is replaced by $f^{1:K}_{t_1:t_2}$. It is well known that in general statistical models may improve greatly when structural changes, even mild though well-aimed, are made to model relevant statistical dependencies, or pruning unimportant and wasteful ones.
Just to illustrate the initial step of the proposed method, let consider a very simple example. Based on common linguistic knowledge let suppose that the LCPM for an application requires only two linguistic features, the thematic tag (sports, fruits, etc.) and the gender inflection (masculine, feminine or neuter) associated to any word, corresponding respectively to the factors f^1_t and f^2_t for the present word, ω_t. The goal is to build a model able to deliver good estimates for $P(f^{1:2}_t|f^{1:2}_{t-M+1:t-1})$. If these features were mutually independent then a simple factorization would lead to $\prod_{i=1}^{2} P(f^i_t|f^i_{t-M+1:t-1})$ to compute these estimates. Let now make the two following assumptions (very reasonable in the Portuguese language): 1) the outcome of f^1_t is conditionally independent of $f^2_{1:t}$, given its own history, i.e., $f^1_t \perp\!\!\!\perp f^2_{1:t}|f^1_{1:t-1}$; and 2) f^2_t depends strongly of f^1_t, even knowing its own history, i.e., $f^2_t \not\perp\!\!\!\perp f^1_{1:t}|f^2_{1:t-1}$. Now, a statistical structure corresponding to $P(f^1_t|f^1_{t-M+1:t-1})P(f^2_t|f^1_{t-M+1:t}, f^2_{t-M+1:t-1})$ should be considered to compute the intended estimates. Such as just illustrated, the method's initial step (formalized in section 3.2) consists of selecting manually the linguistic features at the basis

of the set of factors and establishing a baseline structure.
Prolonging the example above, in order to illustrate the second step of the method, let suppose that the data had shown using appropriate measures that the gender instantiated two words preceding the present word (ω_t) is dominant to help predicting the gender of ω_t, when the theme of ω_t is known. In that case, $P(f_t^2|f_{t-M+1:t}^1, f_{t-M+1:t-1}^{1:2}) \approx P(f_t^2|f_t^1, f_{t-2}^2)$ seems a good approximation. Indeed, the method's second step, which is performed automatically based on a data-driven approach, selects criteriously factors corresponding to past instantiations of the previously selected features and establishes the definitive structure of the model, ultimately pursuing a good compromise between descriptive ability and robustness. The selection criterion essentially uses an information utility measure[2], applying it to the candidate factors in different contexts, such as explained in section3.3, which may bring special advantages in some applications.

3.2 The baseline structure

Jointly, the selected linguistic features must satisfy two main requisites: 1) to convey information that effectively contributes to predict correctly the words in the sentence; and 2) the available data resources fit up the requirements to get robust models. The linguistic features are selected based on common linguistic knowledge relevant for the application, which in general is a relatively simple task that yields a suitable set $f^{1:K}$ [1]. Having in mind the need to achieve a good statistical structure, it follows a procedure to split $f^{1:K}$ into two subsets based on the assumption that some features are conditionally independent of the other ones given its own history. For instance, in the illustrative example in the previous section, $f_t^1 \perp\!\!\!\perp f_{1:t}^2 | f_{1:t-1}^1$ but the conditional independence assumption does not verifies in relation to f_t^2, i.e., $f_t^2 \not\perp\!\!\!\perp f_{1:t}^1 | f_{1:t-1}^2$. This splitting operation is performed non-automatically, once again common linguistic knowledge is in general sufficient to achieve the intended result (in this work was not developed an automatic data-driven method to split $f^{1:K}$, but that is very well feasible). The following conventions are used hereafter: it is assumed that any feature in $f^{1:J}$, with $J < K$, is conditionally independent of any other feature, present or past instantiations, given its own history, i.e., $f_t^i \perp\!\!\!\perp f_{1:t}^j | f_{1:t-1}^i,\ \forall i \neq j,\ 1 \leq i \leq J,\ 1 \leq j \leq K$; and $f^{J+1:K}$ correspond to the remaining features, not satisfying the conditional independence assumption. Accordingly, the baseline LCPM computes the estimates:

$$P(c_t|c_{t-M+1:t-1}) \approx P(f_t^{J+1:K}|f_t^{1:J}, f_{t-M+1:t-1}^{1:K}) \prod_{i=1}^{J} P(f_t^i|f_{t-M+1:t-1}^i) \qquad (4)$$

The product-operator factors can be computed by standard N-grams. The conditional probability corresponding to the "non-independent" features is addressed in the next section.

[1] The lighter notation $f^{1:K}$ is used to express the features set $\{f^1, f^2, \ldots, f^K\}$. The same convention is used, from now on, with $f_{i:j}^{m:n}$ representing a factors set (where $f_\tau^\nu, \quad \tau = i, \ldots, j \quad \nu = m, \ldots, n$ represents the factor corresponding to the linguistic feature f^ν at time τ).

3.3 The structure optimization

In general it is not trivial to train robustly a model able to generate accurate estimates for $P(f_t^{J+1:K}|f_t^{1:J}, f_{t-M+1:t-1}^{1:K})$. Even if M is only a few units and J and K are in the order of the dozens, a fully connected statistical structure is not practicable. Some structural optimization, based on proper criteria, is essential. Follows the presentation of an automatic method that selects only the factors in $f_{t-M+1:t-1}^{1:K}$ satisfying a criteria adapted from the work[1] that lead to the so called Buried Markov Models. In order to simplify the exposition, let introduce the following notation: X, Y and W stand for the sets $f_t^{1:J}$, $f_t^{J+1:K}$ and $f_{t-M+1:t-1}^{1:K}$, respectively; and Z denotes a subset of W. Accordingly, the goal is to find $Z \subset W$ such that robust estimates $P(Y|X,Z)$ approximate well enough $P(Y|X,W)$. Using an Information Theory formulation, any factor $f_\tau^\nu \in W$ candidate to be an element of Z must be selected only if it conveys new information, not provided by those factors already selected or by X, i.e., it must exhibit high score for the conditional mutual information (CMI) $I(Y; f_\tau^\nu|X, Z \setminus f_\tau^\nu)$. This criterion should lead to $|Z|$ factors that, as a whole, exhibit the larger score for the CMI $I(Y; Z|X)$ measured on a sufficiently large and representative data set, so reinforcing the model descriptive power[4]. An extended criterion was introduced in order to favor the selection of factors that increase the difference between the CMI scores measured in different contexts established by X.

Before formalizing the method, the following example illustrates the idea. Keeping the example as simple as possible, let consider that both random variables X and Y are scalars (each represents a single feature): $X \in \{F, S\}$ and $Y \in \{A, B, U\}$. Let suppose that $Y = U$ corresponds to "undefined" category (or simply means unlabeled data) and let also admit that this value, U, brings very little information to the LCPM. Let confront two possible sets for the variable Z, also scalar: $Z^{(1)} \in \{C, D, V\}$ and $Z^{(2)} \in \{E, F, W\}$. In the performed simulation $P(X = F) = 0.6$ (so $P(X = S) = 0.4$) and the conditionals $P(y, z^{(i)}|x),\ i = 1, 2$ are shown in figure 1. According to the criterion referred above, $Z^{(1)}$ is selected instead of $Z^{(2)}$ ($I(Y; Z^{(1)}|X) = 0.177 > I(Y; Z^{(2)}|X) = 0.009$). Indeed, the results in the figure 1 show that in both contexts, $X = F$ or $X = S$, $Z^{(1)}$ is clearly more informative than $Z^{(2)}$ about the outcome of Y. Let suppose now that another data set is used. Running again the simulation are obtained the distributions shown in figure 2. Applying the same criterion as above, now $Z^{(2)}$ is selected instead of $Z^{(1)}$ ($I(Y; Z^{(1)}|X) = 0.208 < I(Y; Z^{(2)}|X) = 0.471$). At first, this result seems acceptable, given the peak $P(y = U, z^{(2)} = W|X)$, on both contexts of X, which does not happens in the case of $Z^{(1)}$. But considering the supposition made that $Y = U$ brings very little information to the LCPM, then this result becomes very unfortunate, since with any of the data sets $Z^{(1)}$ is much more informative than $Z^{(2)}$ about the outcome of Y if not considering $Y = U$. A very interesting evidence provided by the figure 2 is that in the case of $Z^{(2)}$ the results are very similar when comparing both contexts, $X = F$ or $X = S$. And very important too, that similarity does not happens at all in the case of $Z^{(1)}$. It worth's to notice that the same evidence is provided by the results in the figure 1. This illustrative example suggests that a selection criterion based on some measure able to account for the CMI scores estimated in different contexts established by the variable X, should be considered.

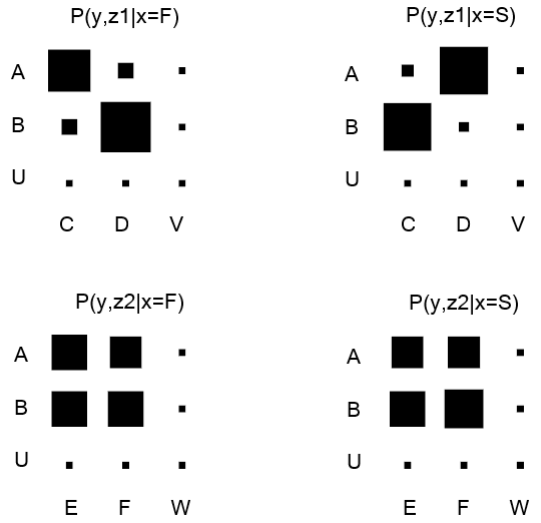

Fig. 1. Probabilistic distributions corresponding to the first data set (Y and Z variables in the vertical and horizontal axis, respectively).

Let begin invoking the cross-context conditional mutual information (CCCMI)

$$I_{X_m}(Y;Z|X=X_n)=\sum_Y\sum_Z P(Y,Z|X_m)\log\frac{P(Y,Z|X_n)}{P(Y|X_n)P(Z|X_n)} \tag{5}$$

If $X_m = X_n$ then obviously results the CMI. The Weighted Utility (WU) measure[1] is defined as follows

$$\begin{aligned} M_{(\lambda)}(Y;Z|X=X_n) &= I(Y;Z|X=X_n) - \\ &-\lambda\sum_{X_m\neq X_n} P(X_m)I_{X_m}(Y;Z|X=X_n) \end{aligned} \tag{6}$$

where $\lambda\in[0,1]$. This measure could be used to implement a criterion so that new components of Z should increase the difference between the CMI and some fraction of the CCCMI average. Finally, let introduce the Global Weighted Utility (GWU) measure[1], that averages the WU based on the distribution of the variable X.

$$N_{(\lambda)}(Y;Z|X)=\sum_{X_m}P(X_m)M_{(\lambda)}(Y;Z|X=X_m) \tag{7}$$

Revisiting the illustrative example above, using the GWU measure with $\lambda=1$, now for both data sets the variable $Z^{(1)}$ is selected instead of $Z^{(2)}$. In the case o the second data set the GWU scores are: $N_{(1.0)}(Y;Z^1|X)=0.427>N_{(1.0)}(Y;Z^2|X)=0.223$. A secure margin separates the scores for the comparing variables, as a consequence of

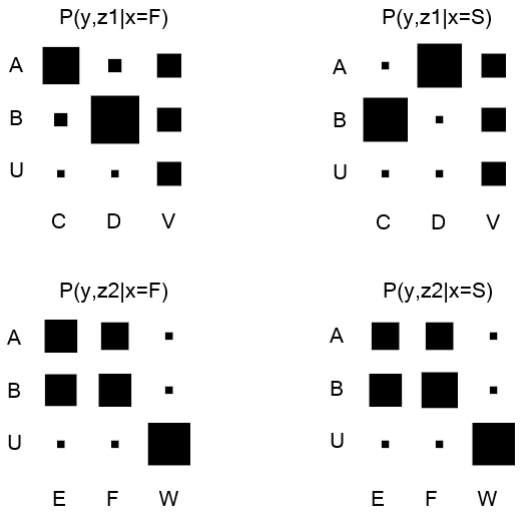

Fig. 2. Probabilistic distributions corresponding to the second data set (Y and Z variables in the vertical and horizontal axis, respectively).

the ability of the GWU measure to capture the variables informativeness differences depending on the context (established by X).
Using the GWU measure, the selection of certain number of factors in W should be relatively straight. Often though the available data is scarce in relation to the dimensions of X, Y and Z, preventing reliable estimates of the defined measures. The proposed algorithm (see "Algorithm 1") follows an iterative approach, eventually finding a sub-optimal solution though more reliable and surely less costly[2]. The strategy is simple, begin with an empty Z and, at each new iteration, add criteriously a component to it. The parameters γ and η must be tuned empirically. Lines 1 to 5 eliminate from the initial candidates set, the linguistic factors that do not convey enough information, using as threshold some fraction of the entropy associated to $f_t^{J+1:K}$ conditioned on $f_t^{1:J}$. Line 6 sorts the remaining factors, placing those with higher GWU scores on the top. Lines 7 to 13 begin with an empty Z, then at each new iteration the factor at the top of the queue sorted in 6 is pulled out and is added to Z if is not redundant in relation to the factors already selected. The process stops when the required dimension of Z, D_z, is reached.

4 Results

The results here presented were obtained from two experiments, kept as simple as possible, though still allowing to en-light key aspects of the proposed method. The data used, extracted from the corpus "CETEMPublico"[10], has a vocabulary with

Algorithm 1 Factors selection (definition of Z)

Require: $f_{t-M+1:t}^{1:K}$, J, K, M, D_Z, λ, γ, η,$Data$
Ensure: Structure of the vector Z
for each $z \in f_{t-M+1:t-1}^{1:K}$ **do**
  if $I(f_t^{J+1:K}; z|f_t^{1:J}) < \gamma H(f_t^{J+1:K}|f_t^{1:J})$ **then**
    Remove z from $f_{t-M+1:t-1}^{1:K}$
  end if
end for
Sort $f_{t-M+1:t-1}^{1:K}$ by descending order of $N_{(\lambda)}(f_t^{J+1:K}; z|f_t^{1:J})$
$Z \Leftarrow \emptyset$
repeat
  $z \Leftarrow$ next non-processed element in $f_{t-M+1:t-1}^{1:K}$
  if $I(f_t^{J+1:K}; z|f_t^{1:J}) > \eta I(z; r|f_t^{1:J}), \forall r \in Z$ **then**
    Add z to Z
  end if
until $|Z| = D_Z$ or all elements of $f_{t-M+1:t-1}^{1:K}$ are processed
return Z

$200K$ words. Just $K = 3$ linguistic features were used, which factors span a window with length $M = 3$. The available data allowed the robust estimation of the used measures. The linguistic factors associated to each word are the morpho-syntactic tag (or *part-of-speech*)[9], the gender inflection and the number inflection, denoted respectively by the symbols m, g and n. The features can take the values: $m_t \in \{\texttt{ADJ}, \texttt{ADV}, \texttt{CONJ}, \texttt{DET}, \texttt{NOM}, \texttt{P}, \texttt{PR}, \texttt{PRP}, \texttt{PRP+DET}, \texttt{V}, \texttt{other}\}$; $g \in G2$ or $g \in G3$, where $G2 = \{\texttt{MASC}, \texttt{FEM}\}$ and $G3 = \{\texttt{MASC}, \texttt{FEM}, \texttt{NEUT}\}$; $n \in G2$ or $n \in G3$, where $N2 = \{\texttt{SING}, \texttt{PLUR}\}$ and $N3 = \{\texttt{SING}, \texttt{PLUR}, \texttt{UNDEF}\}$. Only the feature m is assumed to be independent of the other ones, so $J = 1$ and $f_{t-2:t}^1 = m_{t-2:t}$. Likewise, $f_{t-2:t}^2 = g_{t-2:t}$ and $f_{t-2:t}^3 = n_{t-2:t}$. In order to make clear the results from the two experiments, an adaptation of the algorithm presented in section 3.3 was used, computing separately the factors set Z for each factor in $f_t^{J+1:K}$. Consequently, the joint conditional in equation 4 is approximated considering the factors in $f_t^{J+1:K}$ conditionally independent each other, which does not harms the conclusions of this experiment. Let denote by Z_g and Z_n the subsets of W corresponding, respectively, to the features g and n.

In the initial experiment $g \in G2$ and $n \in N2$. Table 1 presents the factors selection ranking for g and n. In the case of g, such as expected both for $\lambda = 0$ (GWU parameter) or $\lambda = 1$, g_{t-1} is the most informative factor concerning the outcome of g_t. The following rank positions are different depending on λ, but must be noticed that the three first places coincide, g_{t-2} and m_{t-1} are selected the most relevant factors after g_{t-1}. These results seem very reasonable, such as those referring to n, with the past instantiations of n, and then of m and g, by this order, selected as relevant cues for n_t.

In the second experiment $g \in G3$ and $n \in N3$, i.e., are also considered the "neuter" and "undefined" categories for g and n, respectively. Table 2 presents the factors selection ranking for g and n. The results are very surprising, either for n or g, when $\lambda = 0$: n_t is selected as the most relevant factor to inform about the g_t outcome, and

Table 1. Results of experiment 1.

Rank	Z_g		Z_n	
	$\lambda = 0$	$\lambda = 1$	$\lambda = 0$	$\lambda = 1$
1	g_{t-1}	g_{t-1}	n_{t-1}	n_{t-1}
2	g_{t-2}	m_{t-1}	n_{t-2}	n_{t-2}
3	m_{t-1}	g_{t-2}	m_{t-1}	m_{t-1}
4	n_t	m_{t-2}	m_{t-2}	m_{t-2}
5	m_{t-2}	n_{t-1}	g_t	g_{t-1}
6	n_{t-1}	n_t	g_{t-1}	g_{t-2}
7	n_{t-2}	n_{t-2}	g_{t-2}	g_t

Table 2. Results of experiment 2.

Rank	Z_g		Z_n	
	$\lambda = 0$	$\lambda = 1$	$\lambda = 0$	$\lambda = 1$
1	n_t	g_{t-1}	g_t	n_{t-1}
2	g_{t-1}	m_{t-2}	n_{t-1}	m_{t-2}
3	m_{t-1}	g_{t-2}	m_{t-1}	n_{t-2}
4	m_{t-2}	n_{t-2}	n_{t-2}	g_{t-2}
5	g_{t-2}	n_{t-1}	m_{t-2}	g_{t-1}
6	n_{t-1}	m_{t-1}	g_{t-2}	m_{t-1}
7	n_{t-2}	n_t	g_{t-1}	g_t

vice-verso! Only in the second place stand the expected choices: g_{t-1} for g_t and n_{t-1} for n_t. Contrarily, when $\lambda = 1$ the factors selected as the most relevant are precisely the expected ones. It is important to find the explanation for these differences, depending on the value of λ. In large part the explanation is the following. Even for categories of m having no gender inflection, such as verbs for instance, in many data samples of the second data set, $g = \texttt{NEUT}$ and $n = \texttt{UNDEF}$, therefore g and n become quite informative each other (similarly to the illustrative example presented in section 3.3) and consequently when $\lambda = 0$ are obtained unexpected results. When $\lambda = 1$, the GWU measure is able to capture the g and n informativeness differences depending on the context established by m, leading to selections that agree with basic linguistic knowledge. It worth's to emphasize the ability of the method to circumvent this "unfavorable" annotation circumstance, which is not uncommon. Eventually, the method is also able to deal properly with some other "flaws" affecting the corpus. Such as pointed out before, just preliminary experiments were already performed. Further experiments have been planed for comparing the method with other approaches on standard tasks. Nevertheless, the obtained results afford confidence on the method's ability to learn, from the data, good statistical structures for the LCPM in practical applications.

5 Conclusions

In this paper was presented a method for statistical language modeling, designed for an application that satisfies the following conditions: 1) the vocabulary is large, typically at least a few hundred thousand words, in general making difficult to build accurate and robust models; 2) the redundancy patterns inherent to the application can be exploited more efficiently selecting, based on proper criteria and common linguistic knowledge, some set of linguistic features that can be associated to the vocabulary words, following an approach such as the class-dependent N-grams; 3) at least part of these features cannot be modeled independently and the data resources are too scarce to allow building a robust linguistic classes prediction model (LCPM) with fully connected structure. The designed method deals precisely with the problem of optimizing the LCPM structure (the implementation and training problems are not addressed), and complies with two general principles: 1) the overall performance of a statistical model is strongly related to the ability of its structure to represent the application's real dependencies; 2) parsimony favors structures modeling just the relevant statistical dependencies according to some appropriate criterion. The proposed method follows a semi-automatic structure learning approach: after the basic structure being set manually, then a data-driven algorithm, using Information Theory measures, refines the model structure. Although only preliminary experiments were performed, the obtained results show that the method is able to deliver LCPM structures that represent the application's real dependencies and also favor the robustness requirement. Besides, the method presents a remarkable ability to deal with some unfavorable circumstances, or even some flaws, which are not uncommon, affecting the annotation information of the data used to build the models.

References

1. Bilmes, J.: Natural statistical models for automatic speech recognition. In: PhD Thesis. Intl. Computer Science Institute, Berkley, Cal (Oct 1999)
2. Bilmes, J.: Natural statistical models for automatic speech recognition. In: Tech. Report. International Computer Science Institute (Oct 1999)
3. Federico, M.: Tutorial on language models. FBK-irst, Trento, Italy (2010)
4. Hanchuan Peng, F.L., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No.8 (Aug 2005)
5. J. Bilmes, K.K.: Factored language models and generalized parallel backoff. In: Proceedings of HLT/NAACL. pp. 4–6 (2003)
6. K. Kirchhoff, J.B., Duh, K.: Factored language model tutorial. In: Tech. Report. Dept. Electrical Engineering, Univ. of Washington (Feb 2008)
7. Kirchooff, K., Yang, M.: Improved language modeling for statistical machine translation (2005)
8. M. Federico, M.C.: Efficient handling of n-gram language models for statistical machine translation. In: ACL 2007 Workshop on Statistical Machine Translation. pp. 88–95 (2007)
9. Maria Helena Mateus, Ana Maria Brito, I.D., Faria, I.H.: In: Gramática da Língua Portuguesa. Editorial Caminho (1999)
10. Santos, D., Rocha, P.: Evaluating cetempúblico, a free resource for portuguese. In: 39th Annual Meeting of the Association for Computational Linguistics. pp. 442–449 (Jul 2001)