A New Method for Language Modeling

Technical Report TR-ABCP1-14-01 (m+)

Vitor M M C Pera

FEUP - Porto
July 2014

Abstract

This report presents a new method, to the best of the author's knowledge, for statistical language modeling. The method defines the structure of the Language Model based on class dependent N-grams, supporting multiple linguistic classes per word terminal, and on the Factorial Language Model statistical framework. This method differs from formerly published ones essentially in the linguistic classes prediction model, which is built combining in a novel manner a knowledge based approach with a data-driven technique. Based on linguistic expertise it is assumed that only part of the manually selected linguistic factors are conditionally independent of the other ones given their own *histories*. This assumption, which reflects the rôle the factors play in many real applications, leads to a proper factorization and to the partial definition of the statistical model main structure. Then, a data-driven technique is applied for selecting just the factors (linguistic features) required to build an accurate enough model. The proposed algorithm, which is similar to one that has been developed for the so-called Buried Markov Models, may be particularly effective in some circumstance, for instance when some annotation deficiencies affect the text corpus. The effectiveness of this method could be empirically verified in some applications. One experiment succeeded verifying a special ability to select proper factors in some particular circumstances.
....
And based on comparative (in relation to standard class-dependent N-gram) experiments ...
....

Contents

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 that LM.

In the automatic speech recognition (ASR) field in particular, it is well known how strongly the error rates depend on the quality of the LM. The most common approach to ASR language modeling is based on N-grams, typically implementing *bigrams* or *trigrams*, though higher order models, such as *pentagrams*, have been used too. Generally the N-grams are built following a pure data-driven approach, simply based on the relative frequencies of the words sequences observed in some available data. This approach practically avoids any expertise on linguistics, which usually is regarded as very convenient. By the other side, if the vocabulary size exceeds just a few thousands different words, then the sparse data estimation usually becomes a serious problem. Over-fitting occurs and therefore the generalization ability is poor. For instance, if around ten thousand different words exist then even admitting a poorly connected syntax billiards of tri-gram entries (upper bound 10^{12}) must be properly trained, so demanding an huge amount of data containing several repetitions for each tri-gram entry. Nowadays, recognition systems for applications based on vocabularies with hundreds of thousands words are quite common, and obviously then it is not possible to train robustly basic N-grams even if the value of N is kept small.

Statistical modelling techniques based on data sharing or smoothing principles have been developed and are generally used. That is the case of the linear combination of different order N-grams (deleted-interpolation methods)[S. Chen (1998)]. Other common approaches are the N-grams based on back-off strategies and on discount techniques[S. Chen (1996)]. In spite of the inestimable value of those techniques, at least in some specific circumstances, different solutions, for instance directly involving linguistic knowledge, would reach better compromises to mitigate the data sparsity estimation problem.

Consequently, other conceptually different approaches for improving the models have been developed. Some in particular, based on the idea of encoding into the N-gram linguistic knowledge by making explicit relevant linguistic features and their inter-dependencies, have been quite successful. For instance, the so-called class-dependent N-grams[Brown et al.(1992)] have demonstrated to alleviate substantially the data sparsity estimation problem.

In the case of some applications exhibiting relatively complex linguistic patterns involving multiple linguistic features, besides the words, new and better approaches to exploit those patterns are needed[K. Kirchhoff (2008)]. The inefficiencies of the standard class-dependent N-grams that can be used in systems built for such applications are largely due to shortcomings of the models statistical structure, rather than the implementation and training aspects. Hopefully, methods that combine properly linguistic expertise with data-driven structure learning approaches, based on both descriptive and locally discriminative criteria, may improve the models and alleviate the data scarcity effects in those specific applications. Briefly, that was the main motivation for the work here reported.

The structure of this document is as follows. Section 2 presents the class-dependent N-gram based structure supporting the proposed model, considering a variation that does not obligates any word to belong to an unique class. A brief discussion is then presented comparing, in terms of modelling accuracy, this model with a standard N-gram. A few notes concerning the training robustness are also included. Section 3 is dedicated to the linguistic classes prediction model. Initially are addressed specific properties of the applications that motivated the proposed solutions. Then is presented the knowledge based approach that leads to the model main factorization. Finally is presented the data-driven method for selecting the factors used in the classes prediction model. Empirical results obtained using this method are presented in Section 4. The first results are related to the factors selection method special efficiency when, for instance, some specific circumstances affect the annotation data used on training. And in the second part are shown results that were obtained comparing, based on a typical application, the proposed method with the standard class-dependent N-gram based approach.

2 The language model main structure

2.1 The class-dependent N-gram based model

Given a sequence of words $\omega_{1:T}$ [1], the Language Model estimates the probability $P(\omega_{1:T})$. Using the chain rule this probability can be factorized as $\prod_t P(\omega_t|\omega_{1:t-1})$. Let begin assuming that each word, ω', in the vocabulary belongs to a linguistic class, c', (often c' is a multivariate instantiation, representing multiple linguistic features). Then, the conditional probability $P(\omega_t|\omega_{1:t-1})$ could be computed based on the result $P(\omega_t|\omega_{1:t-1}) = P(\omega_t|c_t, \omega_{1:t-1})P(c_t|\omega_{1:t-1})$. Generally, these estimates need to be approximated leading to the common formulation, based on the class-dependent N-gram, $P(\omega_t|\omega_{1:t-1}) \simeq P(\omega_t|c_t, \omega_{t-N+1:t-1})P(c_t|c_{t-M+1:t-1})$.

At this point, it is important to notice that at least in the case of the Portuguese language the assumption made that each word in the vocabulary

[1] Whenever not ambiguous, $\omega_{a:b}$ is used to represent, depending on the context, the sequence of words $\{\omega_a, \omega_{a+1}, \ldots, \omega_b\}$ or the vectorial variable $(\omega_a, \omega_{a+1}, \ldots, \omega_b)$.

belongs to an unique linguistic class is quite abusive in many circumstances. For instance, in the case of using *part-of-speech* tags, some words may be an adverb or else an adjective or a noun, depending on the context, and the same occurs in relation to nouns and some verbal forms, etc.. The existence of many words belonging to several thematic (or semantic) tags, which often is an useful linguistic feature, is just another illustrative example. Therefore, the following model is here used to estimate the conditional probabilities: $P(\omega_t|\omega_{1:t-1}) = \sum_{c_t} P(\omega_t|c_t, \omega_{1:t-1})P(c_t|\omega_{1:t-1})$. Naturally, the summation can be restricted to just a few parcels (for instance using some simple word-dependent indexing mechanism), optimizing the model operation. Such as in the initial model, due to the training data limitations the conditional probabilities in the summation usually need to be approximated by lower order models. The key point is that in general the model $P(\omega_t|c_t, \omega_{1:t-1})$ needs to be more severely simplified than the model $P(c_t|\omega_{1:t-1})$. This is related to the fact that the size of the words vocabulary is much larger than the number of linguistic classes. Let assume that the approximation $P(\omega_t|c_t, \omega_{1:t-1}) \simeq P(\omega_t|c_t, \omega_{t-N+1:t-1})$ is acceptable, corresponding to a class dependent N-gram. In relation to the linguistic classes model, let assume the conditional independence $C_t \perp\!\!\!\perp W_{1:t-1}|C_{t-M+1:t-1}$, that is, it is considered that the previous words do not help significantly to predict C_t if $C_{t-M+1:t-1}$ is known, with M large enough. Therefore, it is accepted the approximation $P(c_t|\omega_{1:t-1}) \simeq P(c_t|c_{t-M+1:t-1})$. Accordingly, the proposed model becomes:

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

Often at least a few tens of thousands different words exist and the available data imposes a very low value for N (for instance, *tri-grams* or even *bi-grams* are typical in ASR). Contrarily, the number of (composite) linguistic categories in general is much lower, let say not exceeding one thousand and, in many applications, well bellow this value. Besides, usually the probabilistic distributions correspondent to the linguistic features are relatively application independent, allowing the use of out of application data. Therefore, in general M can be substantially larger than N.

2.2 Basic assessment of the model quality

An essential quality of the model is its descriptive power. In order to simplify the following analysis, let begin considering again that each word in the vocabulary belongs to an unique linguistic class, so that $P(\omega_t|\omega_{1:t-1}) \simeq P(\omega_t|c_t, \omega_{t-N+1:t-1})P(c_t|c_{t-M+1:t-1})$, where $c = c(\omega), \forall \omega$ is known. Then, the conditional probability expected log-value $\mathcal{E} = E[log(P(\omega_t|\omega_{1:t-1}))]$, over a representative data set $\{\omega_{1:T_l}^{(l)}\}_{l=1}^{L}$, can be approximated as follows:

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

The first term in this expression simply is the conditional probability expected log-value if using a standard N-gram. Concerning to the second term, still keeping the assumption that $c = c(\omega)$, it can be written:

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

which shows clearly that it represents the conditional mutual information $I(c_t;c_{t-M+1:t-N}|c_{t-N+1:t-1})$. Therefore, if on average $c_{t-M+1:t-N}$ conveys relevant information concerning the outcome of c_t, not already conveyed by c_{t-N+1}, then this model has substantially more descriptive power than a standard N-gram. And, such as shown, that difference corresponds to the conditional mutual information $I(c_t;c_{t-M+1:t-N}|c_{t-N+1:t-1})$ measured over some representative data set.

It can be expected that the proposed model becomes especially advantageous if: 1) the linguistic classes sequences observed in the real distribution exhibit a low entropy behaviour, and 2) the value of N is not sufficient large to "capture" this (hidden) behaviour, and 3) the value of M can be made large enough to achieve an accurate enough classes prediction model. The initial condition, corresponding to sharp conditional probabilities $P(c_t|c_{1:t-1})$ over the real distribution, can be observed in many applications (for instance, 2 to 3 bit of information sufficed in some results in Section 4 related to an application with approximately 50 different linguistic classes). And the circumstance that N is small is quite common indeed, due (mainly) to the training resources limitations. Finally, in relation to having a "large M" two key aspects must be stressed, both leading to the satisfaction of this goal: one is related to the fact that the number of linguistic categories often is relatively small (comparing with the words vocabulary size, for instance), facilitating the robust training of higher order models (M-grams); the other is based on the fact that generally the linguistic classes distributions are less application dependent than the words (terminals) distributions, and so additional non-specific data sets may be used to train higher order models.

At this point it is important to recall that the assumption $c = c(\omega)$ is not valid in the proposed model. However, since most of the words belong to just a few linguistic classes, besides that many belong to only one linguistic class, it seems reasonable to extend the conclusion on the descriptive power increase.

....

The following results were obtained running a simulation on a typical application (related to the one presented in the Section 4). Shortly, the data set main characteristics (more detailed presented in the Section 4 are: vocabulary size ... estimated entropy; linguistic features ... size; etc.. Figure **??** shows the likelihood of the proposed model (equation ...), for different configurations of N and M ($N \in \{2, 3, 4\}$ and $N < M \in \{3, 4, 5\}$), over the referred data set. Figure **??** shows the (average) likelihood difference, in relation to a standard N-gram ($N \in 2, 3, 4$), and the estimated MI (also for different values of M and N ...

It is particularly interesting to notice ... according/contrarily to ...

....

Another key aspect concerns to the model robustness. An extensive analysis presents some difficulties, preventing its development here, but a few relevant aspects may be briefly emphasized using an "equivalent" N-gram, which obviously has less parameters to adjust, as a comparative model. Two conditional probability tables (CPTs) need to be trained: the CPT1 refers to the class-dependent N-gram; and the CPT2 contains the M-gram entries corresponding to the linguistic classes prediction model. For small or medium vocabularies and many different linguistic categories (let say, exceeding a few hundreds), likely the CPT2 is larger than the CPT1. Though, in the case of large vocabularies (at least some tens of thousands words), actually when the proposed model is potentially more useful, the CPT2 likely is (for typical values of N and M) much smaller than the CPT1. Therefore, let focus on the CPT1 training robustness. In terms of the *average samples per parameter* (ASPP), at first it seems this ratio is much smaller, by a factor L equal to the number of different linguistic classes, in the proposed model. However, in many typical applications almost every word in the vocabulary maps into just a small part of the linguistic categories set (and many words belong to an unique category), therefore leading to a very sparse CPT1. In conclusion, although eventually it can be necessary to train an almost full CPT1, requiring substantially (L times) more training data comparatively to a standard N-gram, in many useful applications approximate ASPP ratios - and so similar expected robustness - can be expected for the proposed model and the N-gram.

3 The linguistic classes prediction model

3.1 The Factorial Language Model based approach

The linguistic classes prediction model, $P(c_t|c_{t-M+1:t-1})$, is based on the Factorial Language Model (FLM) approach, that was first introduced in [J. Bilmes (2002), J. Bilmes (2003)]. The vocabulary words themselves could also have been treated as factors, i.e., the FLM formalism could have been used at the global model level. A different approach was followed in this work, where the factors represent linguistic features but not the words (terminals), also because of the interest to emphasize some aspects presented in the previous section. Accordingly, c_t is viewed as being composed of K factors, $c_t = f_t^{1:K}$, so that [2] $P(c_t|c_{t-M+1:t-1}) = P(f_t^{1:K}|f_{t-M+1:t-1}^{1:K})$.

Such as it is often applied, the FLM approach ignores the dependencies that eventually exist among the factors. A fundamental assumption in this work is that in many real applications better models can be built improving their structure in order to account for relevant statistical inter-dependencies. Let consider a very simple illustrative application where only two factors, f_t^1 and f_t^2, are used: f_t^1 represents the *thematic* tag (`sports|fruits|...`) associated to the current word; and f_t^2 represents the respective *gender inflection* (`masculine|feminine|neuter`). If these factors are considered independent each other then a simple factorization is performed, leading to the result $P(f_t^{1:2}|f_{t-M+1:t-1}^{1:2}) = \prod_{i=1}^{2} P(f_t^i|f_{t-M+1:t-1}^i)$. Let now consider the two following hypothesis: 1) the outcome of f_t^1 is conditionally independent of $f_{1:t}^2$, given its own *history* ($f_t^1 \perp\!\!\!\perp f_{1:t}^2|f_{1:t-1}^1$); and 2) on the contrary, f_t^2 depends strongly of f_t^1, even knowing its own *history*, $f_{1:t-1}^2$. These hypothesis seem quite reasonable, at least in the case of the Portuguese language: for instance if the *theme* is `fruit` then, contrarily to the `sports` context, likely the *gender* presents a strong statistical bias to the `feminine` category regardless of the *gender-history*; and by the other side, for most applications the *theme* is a relatively conservative linguistic feature. Therefore, eventually it should be more appropriate to consider a little more complex statistical structure leading to the following model: $P(f_t^{1:2}|f_{t-M+1:t-1}^{1:2}) = P(f_t^1|f_{t-M+1:t-1}^1)P(f_t^2|f_t^1, f_{t-M+1:t-1}^{1:2})$. Although this is just a toy-problem, differences such as these among factors naturally occur in many real applications. Accordingly, the development of the proposed linguistic classes prediction model presents an initial step that consists of selecting the respective factors (linguistic features). This step, that is carried out non-automatically, based on linguistic expertise and on the knowledge of the application relevant properties, includes distinguishing the independent and the non-independent factors (f_t^1 and f_t^2, respectively, in the illustrative example). The subsection 3.2 addresses this question in more detail.

[2] Whenever not ambiguous,$f_{t:r}^{m:n}$ is used in this text to represent, for any natural numbers $t \leq r$ and $m \leq n$, $(f_t^m, f_t^{m+1}, \ldots, f_t^n, f_{t+1}^m, f_{t+1}^{m+1}, \ldots, f_{t+1}^n, \ldots, f_r^m, f_r^{m+1}, \ldots, f_r^n)$.

These richer structures in general bring additional training difficulties related to the larger contexts that need to be considered. For instance (using the previous example), in principle good estimates for $P(f_t^2|f_t^1, f_{t-M+1:t-1}^{1:2})$ are more difficult to obtain comparatively to $P(f_t^2|f_{t-M+1:t-1}^2)$. Therefore, it is proposed one method for optimizing the model structure, according to some established restrictions and criterion, trying to achieve a good compromise between the descriptive ability and the training robustness. Moreover, some local discriminative ability, that may have global positive impact in some relatively common training circumstances, is also introduced. It can be advanced, in brief, that the method discards the less relevant "past dependencies" when estimating the non-independent factors conditional probabilities. Based again in the previous example, $P(f_t^2|f_t^1, f_{t-M+1:t-1}^{1:2})$ could be approximated, for instance, by $P(f_t^2|f_t^1, f_{t-2}^2)$, if the method could "perceive" that practically only the *gender* instantiation two words before helps to predict the current word *gender*). This final step in the model development process, that is performed in an automatic manner using a data-driven algorithm based on an approach developed for the Buried Markov Models[J. Bilmes (1999)] used in Automatic Speech Recognition, is discussed in the subsection 3.3.

3.2 The factors selection and the basic model

In simple terms, the candidate factors (linguistic features) are those that jointly convey information contributing to predict correctly the words in the sentence, according to the equation 1 (one needs simply to replace c by f, according to the factors notation). In practice, often other aspects need to be considered when selecting the factors. In particular the required data resources are decisive, possibly preventing the use of potentially useful factors.

Different methods have been used in similar selection problems, from pure knowledge-based approaches to data-driven methods. According to the proposed method the factors are selected non-automatically, based on linguistic expertise. Knowledge about the application linguistic properties that are relevant to accomplish this selection procedure, eventually collected by automatic means, is obviously essential. Nevertheless, one interesting aspect is that very often satisfactory solutions can be easily found, based on relatively basic linguistic knowledge, besides some common sense. And naturally in other cases optimized solutions require quite extensive linguistic expertise. Anyhow, when this task is accomplished must be known the K factors, $f_t^{1:K}$, used in the linguistic classes model.

The FLM approach is often applied ignoring the dependencies that eventually exist among the factors (though the initial formulation does not imposes that assumption), which in general leads to relatively simple models. Such as it was already stressed (subsection 3.1), for many real applications a better compromise between accuracy and robustness (and simplicity) may be achieved if modelling the most relevant inter-dependencies that really occur among the selected factors. These dependencies may directly reflect systematic linguistic knowledge associated to the factors, such as when the *gender inflection* can be viewed as an useful attribute of the *noun* but not of the *verb* (in the Portuguese language, at least, the *gender* is necessarily `neuter` for verbs), or else may be based essentially on relevant statistical dependencies simply observed in the data. Accordingly, the proposed method requires splitting the selected factors into two groups: one group is composed of any factor that (criteriously) is considered to be conditionally independent of the other ones, if its own *history* is known; and the remaining factors compose the other group. Such as for the factors selection task, based on some linguistic expertise and general problem solving skills, this grouping operation must be performed non-automatically. In this work it was not made any attempt to create a dedicated automatic method, though it is obviously possible to use some simple, and eventually effective enough, "ad-hoc" approaches. It must be stressed that eventually the differentiation of the factors into these two groups is also greatly motivated by some conditions (restrictions) beyond the linguistic considerations. In particular, the available training resources may be determinant in this context. For instance, choosing the lesser of two evils, if no appropriate data is available maybe is preferable to model one factor in an isolated manner (considering only its dependency on its own history), even if a relevant statistical dependency exists in relation to other factors, than simply not using that particular factor. Independently of the reasons behind the factors differentiation, by the end of this step it must be known, in relation to each previously selected factor, if it is going to be modeled "independently" or not.

For the sake of clarity and not compromising the model generality, let consider that the "independent" factors correspond to the lower upper-indexes. So, it is assumed that each factor $f_t^1, f_t^2, \ldots, f_t^J$, with $J < K$, is conditionally independent of the others given its own *history*, or equivalently, if $1 \leq i \leq J$ then $f_t^i \perp\!\!\!\perp f_{1:t}^j | f_{1:t-1}^i, \forall j \neq i$. Simultaneously, this property is not valid for the remaining factors, therefore each factor $f_t^{J+1}, f_t^{J+2}, \ldots, f_t^K$ is not going to be modeled as conditionally independent of the other factors.

Then, based on the assumptions made, the baseline linguistic classes prediction model can be expressed as follows:

$$P(c_t | c_{t-M+1:t-1}) = P(f_t^{J+1:K} | f_t^{1:J}, f_{t-M+1:t-1}^{1:K}) \prod_{i=1}^{J} P(f_t^i | f_{t-M+1:t-1}^i) \quad (2)$$

The conditional probabilities $P(f_t^i | f_{t-M+1:t-1}^i)$ correspond to standard M-grams. In relation to the model that computes $P(f_t^{J+1:K} | f_t^{1:J}, f_{t-M+1:t-1}^{1:K})$, in the next section is presented an automatic method that tries to optimize its final structure according to an appropriate objective criterion.

3.3 The model structure optimization

If a few dozen factors are used, which is nothing remarkable, then becomes difficult to build a robust model to compute accurately the probabilities $P(f_t^{J+1:K}|f_t^{1:J}, f_{t-M+1:t-1}^{1:K})$ even if M is small (in general M is not larger than a few units and typically J and K are larger). By the other side, the relevance of the factors and the respective *histories* for the accuracy of this model certainly is quite variable, and likely in many applications some components of $(f_t^{1:J}, f_{t-M+1:t-1}^{1:K})$ are practically irrelevant. Accordingly, the main goal is to develop one method for selecting only these most relevant components. The basic ideas, assumptions made, and criteria behind the proposed method and algorithms, are presented next.

It is assumed that only "past" factors, in $f_{t-M+1:t-1}^{1:K}$, may be discarded. In relation to the "current" factors $f_t^{1:J}$, they must be kept as established in the original model. If in some particular application this restriction is not acceptable then the method may be modified, for instance with minor changes only some of the chosen factors in $f_t^{1:J}$ may continue to be imposed. In order to make the exposition more clear, let begin considering that the symbol Z stands for the random vectorial variable which components are selected from $f_{t-M+1:t-1}^{1:K}$. So, the main goal consists of defining an optimal Z, in the sense that a good compromise is achieved between the accuracy and the robustness of the model $P(f_t^{J+1:K}|f_t^{1:J}, Z)$. Another assumption made is that an unique Z is defined, independently of the $f_t^{1:J}$ particular instantiations, simplifying the implementation of the method. However, given some particular application (and data) for which this restriction could lead to a clearly sub-optimal solution, then the method could be modified allowing different Z structures depending on $f_t^{1:J}$.

The criterion initially established for defining Z was based on general principles of modelling descriptive power and modelling parsimony. Therefore, each selected factor should convey information, not already present in $f_t^{1:J}$ or in any other selected factor, effectively relevant to predict correctly the outcome of $f_t^{J+1:K}$. Using an Information Theory common formulation, any factor $f_\tau^\iota \in f_{t-M+1:t-1}^{1:K}$ candidate to be a component of Z should exhibit, over a representative data set, a high score for the conditional mutual information $I(f_t^{J+1:K}; f_\tau^\iota|f_t^{1:J}, f_{t-M+1:t-1}^{1:K} \setminus f_\tau^\iota)$. In the proposed method the size of Z is previously established (parameter $|Z|$), so, this criterion should lead to choosing the $|Z|$ factors that, as a whole, exhibit the larger score for the conditional mutual information $I(f_t^{J+1:K}; Z|f_t^{1:J})$. Or, equivalently, the selected Z should maximize the conditional likelihood $P(f_t^{J+1:K}|f_t^{1:J}, Z)$ over a sufficient large and representative data set, so reinforcing the model descriptive power.

The definitive criterion for selecting the factors was later established, complementing the initial criteria with the idea of favouring the factors that increase, over some representative data set, the difference between the conditional likelihoods in the correct context, established by $f_t^{1:J}$, relatively to the incorrect contexts. This local "discriminative ability" obviously does not extends to the global model, since the linguistic classes, including the $f_t^{1:J}$ factors, are marginalized according to the equations 1 and 2. Nevertheless, in some circumstances involving the training data this ability may have a relevant positive impact on the factors selection (such as it is going to be illustrated next and, using a more concrete example, again in the Section 4). In order to make the exposition more clear, in some of the following expressions a lighter notation is used (whenever the original notation is more appropriate, it continues to be used): the symbol X denotes $f_t^{1:J}$, i.e., X is the random vectorial variable representing the current values of the "independent" factors; and the symbol Y denotes $f_t^{J+1:K}$, i.e., Y is the random vectorial variable representing the current values of the "dependent" factors. Accordingly, the expressions $P(f_t^{J+1:K}|f_t^{1:J}, Z)$ and $P(Y|X;Z)$ have the same meaning, such as $I(f_t^{J+1:K}; Z|f_t^{1:J})$ and $I(Y;Z|X)$ are identical.

Let now consider a simple illustrative variable selection (or structure learning) problem. For the sake of clarity and not compromising the idea, the random variables X, Y and Z are scalars (or unidimensional) and may take just a few different values, i.e., the number of different classes is very small: $X \in \{\mathtt{F}, \mathtt{S}\}$ (for instance, meaning that X is the *theme* `F` or else `S`, respectively, `fruit` or `sport`); $Y \in \{\mathtt{A}, \mathtt{B}, \mathtt{U}\}$ (Y corresponds to another linguistic feature) with the value `U` meaning that Y is `unidentified` or `unknown`; $Z^{(1)} \in \{\mathtt{C}, \mathtt{D}, \mathtt{V}\}$, which is one of the candidates to be selected as Z, with `V` standing for the `unknown` or `unidentified` value); and $Z^{(2)} \in \{\mathtt{E}, \mathtt{F}, \mathtt{W}\}$ is the alternative candidate for Z, with `W` meaning `unknown` or `unidentified`). Let consider that a representative data set presents the conditional probability distributions $P(y, z^{(1)}|x)$ and $P(y, z^{(2)}|x)$ shown in the Figure 1 [3] (due to edition difficulties, in the figure $z^{(n)}$ is typed as zn). In the simulation were used the probabilities $P(X = F) = 0.6$ and $P(X = S) = 0.4$. According to the initial selection criterion, $Z^{(1)}$ is selected since $I(Y; Z^{(1)}|X) = 0.177 > 0.009 = I(Y; Z^{(2)}|X)$. Indeed, the pictures clearly show, in both contexts $X = S$ or $X = F$, that $Z^{(1)}$ is much more informative than $Z^{(2)}$ about the outcome of Y. Now, let consider that more data is collected and added to the initial set, so that the distributions $P(y, z^{(1)}|x)$ and $P(y, z^{(2)}|x)$ become, in this final data set (let name it this way), as shown in the Figure 2. Such as it can be observed, contrarily to the initial data set, a substantial part of the new data belongs to the `unknown/unidentified` classes. And in the case of $Z^{(1)}$ this `unknown/unidentified` data distributes almost uniformly by the Y classes, while in the case of $Z^{(2)}$ it concentrates in the class $Y = \mathtt{U}$ (let notice that the essential final results would continue valid

[3] In `http://speech-rec-vcp.com` is available the Excel file `newLMsml.xls` containing these distributions, such as other used later, and the simulation used to generate the results that are presented next.

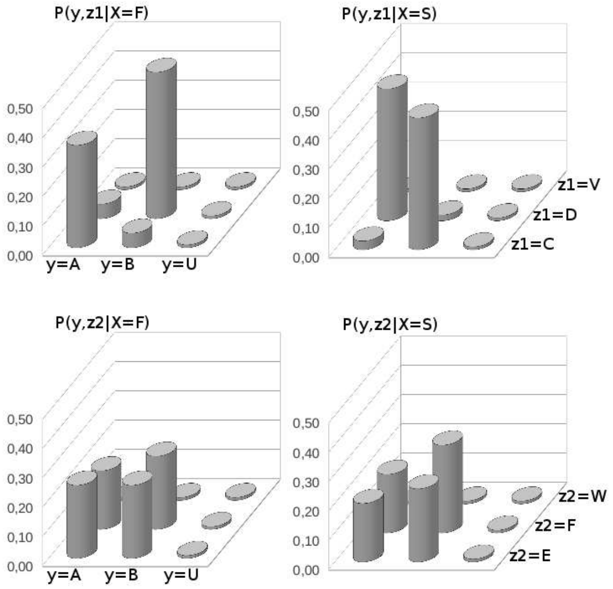

Figure 1: Probabilist distributions corresponding to the initial data set.

if the `unknown/unidentified` $(y, z^{(1)} = V)$ distribution was also focused). Applying again the initial selection criterion, one can verify that now $Z^{(2)}$ is selected, since $I(Y; Z^{(1)}|X) = 0.208 < 0.471 = I(Y; Z^{(2)}|X)$. It is interesting to notice that the distributions corresponding to $Z^{(2)}$, in both contexts $X = F$ or $X = S$, are quite similar. And this fact is specially relevant because it is related to the probabilistic distribution flatness of the "correctly labelled" part of the data, which reflects the circumstance that the factor $Z^{(2)}$ conveys very little information about the Y outcome. This is an essential aspect, and obviously the situation is very different in relation to $Z^{(1)}$. This suggests that these differences, in relation to $Z^{(1)}$ and $Z^{(2)}$, and depending on the contexts established by X, could favour a selection method based on some discriminative parameters adaptation (training), or features selection (or structure learning), approach. Accordingly, the "new" selection criterion and consequent method can be formulated as follows. The cross-context conditional mutual information (CCCMI) is an information theory measure at the basis of this method, since it estimates the mutual information (MI) between Y and Z under the class $X = X_n$, considering the context $X = X_m$:

$$I_{X_m}(Y; Z|X = X_n) = \sum_{Y} \sum_{Z} P(Y, Z|X_m) \log \frac{P(Y, Z|X_n)}{P(Y|X_n)P(Z|X_n)} \qquad (3)$$

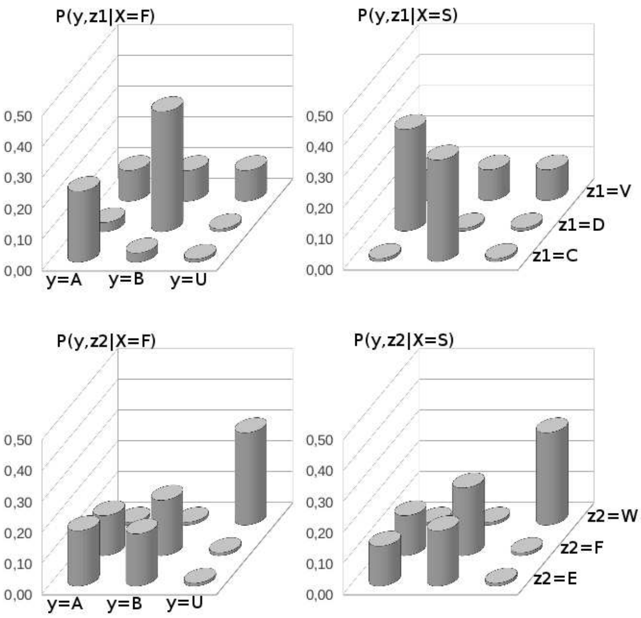

Figure 2: Probabilist distributions corresponding to the final (enlarged) data set.

If $X_m = X_n$ then, obviously, the CCCMI becomes the conditional mutual information (CMI), already denoted simply as $I(Y;Z|X = X_n)$. Using the CMI and the CCCMI it can be defined the following measure

$$\begin{aligned} M_{(\lambda)}(Y;Z|X = X_n) &= I(Y;Z|X = X_n) - \\ &-\lambda \sum_{X_m \neq X_n} P(X_m) I_{X_m}(Y;Z|X = X_n) \qquad (4) \end{aligned}$$

where $\lambda \in [0,1]$. Let name this measure *Weighted Utility* (WU) (regarding to the *Utility* measure in [J. Bilmes (1999)]). If $\lambda \in]0,1]$ then the WU discounts to the CMI measure some fraction of the CCCMI weighted average, according to the priors $P(X)$ (obviously, the WU simply becomes the CMI if $\lambda = 0$). Therefore, for any particular $X = X_n$, this measure could be at the basis of a criterion for selecting the components of Z, in a manner that the dependencies associated to these new components should increase the difference between the CMI (when the context is *correct*) and a CCCMI weighted average when the contexts are *incorrect* ($\forall X \neq X_n$). A new measure must be defined, the *Global Weighted Utility* (GWU) estimating the WU average weighted according to the probabilities $P(X)$

$$N_{(\lambda)}(Y;Z|X) = \sum_{X_n} P(X_n) M_{(\lambda)}(Y;Z|X = X_n) \qquad (5)$$

Obviously, also the GWU allows to adjust the relative strength of the CMI and the CCCMI scores, meaning that it can be favoured the model descriptive ability, leading to a structure that in principle increases the likelihood of the typical data samples, or else favouring a structure that has some potential to locally discriminate better, based on the context established by X. Although this ability does not extends to the global model, by the other side it can be expected that in some circumstances this criterion effectively brings practical benefits, such as the current example can illustrate (and also some more concrete results in the Section 4 show). If applying this measure (GWU), with the parameter $\lambda = 1.0$, to the present illustrative example, with the final (augmented) data set, then $Z^{(1)}$ is selected since $N_{1.0}(Y; Z^{(1)}|X) = 0.427 > 0.223 = N_{1.0}(Y; Z^{(2)}|X)$. Recalling the result obtained with the simpler CMI criterion, $Z^{(2)}$ was then selected because "its" CMI increased much more than the CMI related to $Z^{(1)}$: $\Delta I(Y; Z^{(1)}|X) = 0.208 - 0.177 = 0.031$ and $\Delta I(Y; Z^{(2)}|X) = 0.427 - 0.009 = 0.419$. But that $Z^{(2)}$ increase "spreads" over all the contexts, also very strongly in the "wrong" contexts (practically uniformly, the difference is 0.42 for all CCCMI estimates), such as the Table 1 shows (due to the already emphasized independence related "distribution flatness"), so the "new" criterion is able to prevent choosing $Z^{(2)}$. It worth to notice that the "new" criteria does not change the (optimal) result obtained applying the initial criterion (max CMI) over the initial data set [4].

$I_{\beta}^{initial}(Y; Z\|\alpha)$	$\alpha = \mathtt{F}$	$\alpha = \mathtt{S}$
$\beta = \mathtt{F}$	0.009	0.006
$\beta = \mathtt{S}$	0.009	0.011

$I_{\beta}^{final}(Y; Z\|\alpha)$	$\alpha = \mathtt{F}$	$\alpha = \mathtt{S}$
$\beta = \mathtt{F}$	0.426	0.422
$\beta = \mathtt{S}$	0.426	0.428

Table 1: The CCCMI estimates for the initial and the final data sets ($I_{\beta}^{initial}(Y; Z|\alpha)$ is a lighter notation of $I_{X=\beta}^{initial}(Y; Z|X = \alpha)$, i.e., the CCCMI in relation to the "model" $P(y, z|X = \alpha)$ estimated over the data $\{(y, z, X = \beta)\}$).

The implementation of an automatic method to define the structure of Z would then be straight, choosing the combination of a certain number of factors in $f_{t-M+1:t-1}^{1:K}$ essentially based on the criteria of maximizing the GWU measure, if the needed computational resources were available and enough data could be collected. On the contrary, quite often the data is sparse given the dimension of the vectors X, Y and Z, preventing reliable estimates of the defined measures and, consequently (eventually other reasons exist too), the simultaneous selection of the Z components (factors in $f_{t-M+1:t-1}^{1:K}$) becomes practically impossible. The proposed method follows

[4] The pointed out Excel file `newLMsml.xls` fully contains the simulation ran for this illustrative example.

an iterative approach to find an approximate solution, eventually suboptimal but hopefully more robust and certainly less costly, for this discrete space problem. Accordingly, the Algorithm 1 is based on the simple strategy of beginning with an empty Z and, at each new iteration, adding criteriously a component to it. The Algorithm 1 requires the knowledge of $f_t^{1:J}$, $f_t^{J+1:K}$

Algorithm 1 Factors selection (definition of Z)

```
Require: $f_{t-M+1:t}^{1:K}$, $J$, $K$, $M$, $D_Z$, $\lambda$, $\gamma$, $\eta$,$T_{LF}$
Ensure: Structure of the vector $Z$
 1: for each $z \in f_{t-M+1:t-1}^{1:K}$ do
 2:    if $I(f_t^{J+1:K}; z|f_t^{1:J}) < \gamma H(f_t^{J+1:K}|f_t^{1:J})$ then
 3:       Remove $z$ from $f_{t-M+1:t-1}^{1:K}$
 4:    end if
 5: end for
 6: Sort $f_{t-M+1:t-1}^{1:K}$ by descending order of $N_{(\lambda)}(f_t^{J+1:K}; z|f_t^{1:J})$
 7: $Z \Leftarrow \emptyset$
 8: repeat
 9:    $z \Leftarrow$ next non-processed element in $f_{t-M+1:t-1}^{1:K}$
10:    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
11:       Add $z$ to $Z$
12:    end if
13: until $|Z| = D_Z$ or all elements of $f_{t-M+1:t-1}^{1:K}$ are processed
14: return $Z$
```

and also $f_{t-M+1:t-1}^{1:K}$, the candidates for components of Z (or equivalently, $f_{t-M+1:t}^{1:K}$, J, K and M). The dimension, D_Z, of the vector Z must also be known. Requires also a representative dataset, T_{LF}, with enough samples of the linguistic features associated to $f_{t-M+1:t}^{1:K}$. Besides, the algorithmic parameters λ, γ and η must be set previously and must be known: λ is used in the GWU measure, according to equations 4 and 5; γ is used to access the relevance of each factor (representing a linguistic feature); and η is used to access the redundancy associated to each factor. Essentially, the Algorithm 1 comprises the following steps (the functions represented by the symbols H and I are used, respectively, for the Conditional Entropy and the CMI): 1) eliminate from $f_{t-M+1:t-1}^{1:K}$ the linguistic factors that do not convey enough information about $f_t^{J+1:K}$ conditioned on $f_t^{1:J}$ (lines 1 to 5); 2) continue preparing the iteration process, sorting the remaining factors so that those corresponding to higher GWU scores are considered more relevant and therefore are given higher priority (line 6); 3) initialize Z (line 7); 4) iterate through all the remaining factors, according to the established relevance ranking, selecting only those ones that, based on some heuristic, are not redundant in relation to all the factors already selected (lines 8 to 13).

In the following section are presented results based on a modified version of this algorithm. That modification may be necessary in some applications mainly because of the difficulty to estimate robustly the measures needed to run this algorithm, due to the relatively large number of factors in $f_t^{J+1:K}$ (indeed, in many applications the "dependent" linguistic features can be quite numerous). In those cases, it can be tried to establish Z assuming that the probabilistic joint-distribution associated to the factors in $f_t^{J+1:K}$, conditioned on $f_t^{1:J}$ and $f_{t-M+1:t-1}^{1:K}$, can be fairly approximated considering that those factors are conditionally independent each other. Sometimes this approximation can be too much crude, but in other relevant applications may be acceptable. With this modification, becomes possible to choose specific components of Z for each $f_t^J, f_t^{J+1}, \dots, f_t^K$ and, besides, the dimension of each Z may also be different. Therefore, instead of the main structure corresponding to equation 2, the linguistic classes prediction model becomes

$$P(c_t|c_{t-M+1:t-1}) = \prod_{i=J+1}^{K} P(f_t^i|f_t^{1:J}, f_{t-M+1:t-1}^{1:K}) \prod_{i=1}^{J} P(f_t^i|f_{t-M+1:t-1}^i) \tag{6}$$

The operations performed in the Algorithm 2 are very similar to those in the Algorithm 1, essentially differing on that the selection procedure is repeated, separately, for each factor in $f_t^{J+1:K}$. Obviously, other modifications can be made to improve any of these baseline algorithms in order to deal more properly with some other limitations.

Algorithm 2 Factors selection (definition of $Z_{J+1}, Z_{J+2}, \dots, Z_K$)

```
Require: $f_{t-Nc+1:t}^{1:K}$, $J$, $K$, $N_c$, $D_{Z_i}(\forall i)$, $\lambda$, $\gamma$, $\eta$,$T_{LF}$
Ensure: Structure of the vector $Z_{J+1}, Z_{J+2}, \dots, Z_K$
for each factor $f_t^j \in \{f_t^{J+1j}, f_t^{J+2}, \dots, f_t^K\}$ do
    $F \Leftarrow f_{t-Nc+1:t-1}^{1:K}$
    for each $z \in F$ do
        if $I(f_t^j; z|f_t^{1:J}) < \gamma H(f_t^j|f_t^{1:J})$ then
            Remove $z$ from $F$
        end if
    end for
    Sort $F$ by descending order of $N_{(\lambda)}(f_t^j; z|f_t^{1:J})$
    $Z_j \Leftarrow \emptyset$
    repeat
        $z \Leftarrow$ next non-processed element in $F$
        if $I(f_t^j; z|f_t^{1:J}) > \eta I(z; r|f_t^{1:J}), \forall r \in Z$ then
            Add $z$ to $Z$
        end if
    until $|Z_j| = D_{Z_j}$ or all elements of $F$ are processed
    return $Z_j$
end for
```

4 Results

4.1 Model structure optimization experiment

These results were obtained considering very simple applications of the exposed method, in order to facilitate the experimental work. The vocabulary is large, there are approximately 200K different words, but only three factors are used, spanning along word sequences with length (also) three. Even if this is like a toy-problem, the obtained results illustrate the proposed method and allow to emphasize some important properties.

Also in order to make easy emphasizing a few key aspects, and not compromising the exposition, an unique "isolated" factor is considered (so, $J = 1$). This factor represents the morpho-syntactic tag (the *part-of-speech*) associated to the actual hypothesized word, therefore, using a more intuitive notation, $f_t^1 = m_t$ (generalizing for m_{t-1}, m_{t-2}, etc.). Eleven different classes (including the class `other`) are considered, so that $m_t \in \{\mathtt{ADJ,ADV,CONJ,DET,NOM,P,PR,PRP,PRP+DET,V,other}\}$ [5]. For the same reason, also only two "non-isolated" factors are considered, f_t^2 and f_t^3, representing respectively the *gender inflection* and the *number inflection* associated to the current word. Accordingly, the following notation is going to be used: g_t instead of f_t^2, and n_t replacing f_t^3 (generalizing for g_{t-1}, n_{t-1}, etc.). It is important to emphasize that the choice of these three factors, and their "isolated" *versus* "non-isolated" splitting, was based not only on linguistic considerations but also, among other practical aspects, was conditioned by the available annotation material (for the 200K words *corpus*).

Table 4.1 shows the notation used to represent the factors associated to the (length 3) words sequences, and its correspondence in terms of linguistic features and word position (symbol t). The initial results that are going to be

Symbol	Linguistic feature	Position
m_t	morpho-syntaxe tag	actual word
m_{t-1}	morpho-syntaxe tag	previous word
m_{t-2}	morpho-syntaxe tag	two words before
g_t	gender inflection	actual word
g_{t-1}	gender inflection	previous word
g_{t-2}	gender inflection	two words before
n_t	number inflection	actual word
n_{t-1}	number inflection	previous word
n_{t-2}	number inflection	two words before

Table 2: Notation and meaning associated to the factors used.

[5] Further details, including the program used to create those *part-of-speech* tags, can be found in `http://speech-rec-vcp.com` .

presented were obtained using only words which factor m_t can be associated to the factor g_t taking the value `masculine`, or else `feminine`; for instance, in the Portuguese language, adjectives (`ADJ`) are included but verbs (`V`) are not. So, these results consider $g_t \in G2 = \{\texttt{masculine}, \texttt{feminine}\}$. To allow some important remarks, later results consider also the words corresponding to a third class, `neuter`, for the *gender inflection*, either because the related principal feature has no *gender inflection* (e.g., the verbs), or because that information is not available (indeed, these results used the whole available data, which is incompletely annotated). Accordingly, these results consider $g_t \in G3 = \{\texttt{masculine}, \texttt{feminine}, \texttt{unknown}\}$. In an analogous manner, in the initial results the *number inflection* considers just two classes, so that $n_t \in N2 = \{\texttt{singular}, \texttt{plural}\}$, and for the later results is also considered a third class, so that $n_t \in N3 = \{\texttt{singular}, \texttt{plural}, \texttt{unknown}\}$.

Such as noted before, all these results were obtained using the modified version (Algorithm 2, in the Appendix) of the proposed algorithm. Therefore, specific factors Z_g and Z_n are defined for g_t and n_t, respectively.

The results in Table 2 were obtained with $g_t \in G2$. Automatically, the vector Z_g is defined as the table shows, for the two extreme values of the λ parameter. It can be seen that if only one factor is selected ($D_{Z_g} = 1$), then for the two extreme values of λ the same feature is selected, which is, such as expected, the *gender inflection* associated to the previous word. The following factors have their selection order depending on λ (it could be confirmed that in the case of values between 0 and 1 those differences occur mostly in "the last choices" if λ is small, such as expected). However, for instance if $D_{Z_g} = 3$ (three factors are selected), the same features are proposed for both values of λ (the second and third choices change places). For larger values of D_{Z_g} much less informative features are available, either considering only the correct contexts or discounting the contributions for the incorrect contexts, and the differences on the selection order also increase. However, in practice these "last choices" do not deserve any attention.

Order	$\lambda = 0$	$\lambda = 1$
fisrt	g_{t-1}	g_{t-1}
second	g_{t-2}	m_{t-1}
third	m_{t-1}	g_{t-2}
fourth	n_t	m_{t-2}
fifht	m_{t-2}	n_{t-1}
sixth	n_{t-1}	n_t
seventh	n_{t-2}	n_{t-2}

Table 3: Automatic selection order of the factors in Z_g when the *gender inflection* of the actual word $g_t \in G2$ (`masculine` or `feminine`).

The key point is that the agreement scenario (for $g_t \in G2$) for the "first choices" changes substantially if considering $g_t \in G3$, such as the results in the Table 3 show. Now, if $D_{Z_g} = 1$ and $\lambda = 0$ the factor automatically selected is the *number inflection* associated to the actual word, which is a very surprising result. By the other side, if $\lambda = 1$ the factor selected is the *gender inflection* associated to the previous word, which seems a much more reasonable choice. Comparing these results with those obtained when $G = 2$ it can be perceived that the local ability to discriminate (when $\lambda = 1$) allowed to leave out the n_t factor because even if it conveys a relatively large amount of information about the g_t outcome, that occurs for all the contexts (correct or incorrect) established by m_t. It happens that in this experiment, the available data has a especially large effect in this different "first choice", since even for the classes of m_t that have no *gender inflection*, such as the verbs for instance, in many data samples both g_t and n_t take the value `unknown`. Anyhow, this experiment still illustrates the potential benefit due to the hybrid descriptive and (locally) discriminative criterion used to decide the structure of the model $P(f_t^{2:3}|f_t^1, f_{t-2:t-1}^{1:3})$.

Order	$\lambda = 0$	$\lambda = 1$
first	n_t	g_{t-1}
second	g_{t-1}	m_{t-2}
third	m_{t-1}	g_{t-2}
fourth	m_{t-2}	n_{t-2}
fifth	g_{t-2}	n_{t-1}
sixth	n_{t-1}	m_{t-1}
seventh	n_{t-2}	n_t

Table 4: Automatic selection order of the factors in Z_g when the *gender inflection* of actual word, $g_t \in G3$ (`masculine`, `feminine` or `unknown`).

Table 4 shows the results, analogous to those already presented, now in relation to the factor $n_t \in N2$. Automatically, the vector Z_n is defined as the table shows, for the two extreme values of the λ parameter. It can be seen that if $D_{Z_n} \leq 4$, then the same factors are selected, and by the same order, for any of the two extreme values of λ, with the *number inflection* associated to the previous word being the "first choice", such as expected. The differences only occur for the "last choices", which in principle are completely irrelevant.

The results presented in the Table 5 refer to $n_t \in N3$. Now, if $D_{Z_n} = 1$ and $\lambda = 0$ the factor automatically selected is the *gender inflection* associated to the actual word, and if $\lambda = 1$ the feature selected is the *number inflection* associated to the previous word. Once again, "without discrim-

Order	$\lambda = 0$	$\lambda = 1$
first	n_{t-1}	n_{t-1}
second	n_{t-2}	n_{t-2}
third	m_{t-1}	m_{t-1}
fourth	m_{t-2}	m_{t-2}
fifth	g_t	g_{t-1}
sixth	g_{t-1}	g_{t-2}
seventh	g_{t-2}	g_t

Table 5: Automatic selection order of the factors in Z_n when the *number inflection* of the actual word, $n_t \in N2$ (`singular` or `plural`).

Order	$\lambda = 0$	$\lambda = 1$
first	g_t	n_{t-1}
second	n_{t-1}	m_{t-2}
third	m_{t-1}	n_{t-2}
fourth	n_{t-2}	g_{t-2}
fifth	m_{t-2}	g_{t-1}
sixth	g_{t-2}	m_{t-1}
seventh	g_{t-1}	g_t

Table 6: Automatic selection order of the factors in Z_n when the *number inflection* of actual word, $n_t \in N3$ (`singular`, `plural` or `unknown`).

ination" ($\lambda = 0$) the choice is rather surprising and "with discrimination" the choice is reasonable. The explanation for this behaviour is essentially the same of the g_t factor.

Finally, the results in the Table 6 show that a small value for the parameter λ allows to change substantially the factors selection (in this particular case, $\lambda = 0.1$ is enough to "eliminate" g_t as the first candidate), and although these results refer to an artificially simple problem, they still allow to illustrate the (non-surprising) fact that the range $[0, 1]$ can be enough to become effective.

The results presented were obtained with the values of the parameters γ and η in a "high tolerance" range. This means that γ is set so that all factors (linguistic features) are considered to be relevant enough and η is set so that the redundancy criteria becomes very loose too. Obviously in any practical implementation at some moment those parameters should be tuned according to more restrictive selection criteria, but in the scope of this exposition those optimizations are not relevant. Anyhow, all along the many experiments performed, in some cases with these parameters adjusted to strongly restrict less-relevant and/or somewhat redundant features so that just one

$\lambda = 0.0$	$\lambda = 0.1$	$\dots$	$\lambda = 0.4$	$\lambda = 0.5$	$\dots$	$\lambda = 1.0$
g_t	n_{t-1}	$\dots$	n_{t-1}	n_{t-1}	$\dots$	n_{t-1}
n_{t-1}	n_{t-2}	$\dots$	n_{t-2}	m_{t-2}	$\dots$	m_{t-2}
m_{t-1}	m_{t-2}	$\dots$	m_{t-2}	n_{t-2}	$\dots$	n_{t-2}

Table 7: The three selected factor (first choice in the upper line) associated to Z_n, when $n_t \in N3$ and $D_{Z_n} = 3$, for different values of λ (from 0.0 to 1.0 with increments of 0.1).

or two features could survive (it worth to notice that the selection priority is based on the N-measure while the relevance and redundancy criteria are based on the CMI measure), the general behaviour was observed and the descriptive *versus* discriminative effects could be observed.

4.2 Comparing with the baseline class-dependent N-gram

....
...
....

5 Conclusions

....
In this report was presented a new method, to the best of the author's knowledge, for statistical language modeling. The technical problem that the method tries to solve, following a semi-automatic structure learning approach, is, ultimately, alleviating the data sparsity problem. More in concrete, it is assumed that the proposed solution is designed for a LM dedicated to an application, in ASR for instance, satisfying the conditions:

- the vocabulary size is large (typically with more than a few thousand words) given the available training data, so the standard methods do not prevent, at least entirely, the *over-fitting* effects;

- the redundancy patterns inherent to the application can be exploited more efficiently criteriously selecting, based also on linguistic expertise, several linguistic features;

- some of these features can be modeled in an "isolated" manner (e.g., in many applications the *thematic tag* may be properly modeled considering only its own *history*), and other(s) in an "non-isolated" manner (e.g., the *gender inflection* eventually depends significantly on the *thematic tag*) according to the statistical inter-dependencies naturally associated to their behaviour.

Two main ideas are at the basis of the proposed semi-automatic method (that address only the language modeling problem at the structural level, in relation to the implementation and training, well known approaches may be followed). One is related to the general principle that the overall performance of any statistical model is directly related to the incorporation of relevant knowledge about the real application. In this particular problem, the structure of the model, based on the class-dependent N-grams and the FLM formalisms, should be tailored, based on linguistic expertise, according to the real properties of the application. The other idea concerns to the general principle of parsimony, favouring models with compact structure. Accordingly, the proposed method establishes the final structure trying to preserve just the relevant statistical dependencies. The objective criteria used to assess that relevance considers the contributions of those dependencies to the model descriptive power and also to acquire some local discriminative ability.

Although just preliminary experiments were performed, the obtained results seem to confirm the confidence on the method main strengths, ultimately residing on the complementarity achieved combining: 1) linguistic expertise to directly encode, at the structural level, relevant knowledge; 2) an automatic data-driven method based on objective criteria and on solid information theory concepts, trying to achieve a parsimonious structure preserving enough descriptive power and presenting some local discriminative ability that brings interesting properties.
....

References

[J. Bilmes (1999)] J. Bilmes. Tech. Report TR-99-016, "Natural Statistical Models for Automatic Speech Recognition". International Computer Science Institute. Oct. 1999.

[K. Kirchhoff (2008)] K. Kirchhoff, J. Bilmes, and K. Duh. "Factored Language Model Tutorial", Tech. Rep. UWEETR-2008-0004. Dept. Electrical Engineering, Univ. of Washington. Feb. 2008.

[J. Bilmes (2002)] J. Bilmes et al.. Novel speech recognition models for arabic. JHU 2002 summer workshop final report. 2002.

[J. Bilmes (2003)] J. Bilmes, K. Kirchhoff. Factored language models and generalized parallel backoff. In Proceedings of HLT/NAACL, pages 4-6. 2003.

[Brown et al.(1992)] P.F. Brown, V.J. Della Pietra, P.V. DeSouza, J.C. Lai, and R.L. Mercer. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18:467–479, 1992.

[S. Chen (1996)] Stanley F. Chen. Building Probabilistic Models for Natural Language. PhD thesis, Harvard U., 1996.

[S. Chen (1998)] S. F. Chen and J. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Tech. Report TR-10-98, Computer Science Group Harvard U., Cambridge, August 1998 (original postscript document).

[H. Mateus (1999)] Maria Helena Mateus, Ana Maria Brito, Inês Duarte, and Isabel Hub Faria. Gramática da Língua Portuguesa. Editorial Caminho, 1999.

[M. Federico (2010)] Marcello Federico. Tutorial on Language Models. FBK-irst, Trento, Italy, 2010.

[Katrin Kirchhoff (2008)] K. Kirchhoff, J. Bilmes, and K. Duh. Factored Language Models Tutorial, Tech. Report UWEETR-2007-0003, Dept. of EE, U. Washington, 2007.

[H. Schmid(1995)] Helmut Schmid. Improvments in Part-of-Speech Tagging with and Application to German. Proceedings of the ACL SIGDAT-Workshop, pp. 47-50, 1995.

[S. Chen (1996)] Stanley F. Chen, and Joshua T. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proc. of the 34th Annual Meeting of the ACL (ACL '96)*, 1996.

[M. Federico (2010)] M. Federico, N. Bertoldi and M. Cettolo. IRST Language Modeling Toolkit (Version 5.50.02) - User Manual. FBK-irst, Trento, Italy, 2010.

[P. Clarkson (1997)] P. Clarkson and R. Rosenfeld. Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, 1997.

[A. Stolcke (2002)] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. Spoken Language Processing*, Denver, USA, 2002.