On the Use of Morphological Constraints in N-gram Statistical Language Model

A. Ghaoui[(1)(2)(3)], F. Yvon[(2)], C. Mokbel[(1)] et G. Chollet[(2)]

[(1)] University of Balamand BP 100 Tripoli, Lebanon
Chafic.Mokbel@balamand.edu.lb
[(2)] CNRS-URA820, Ecole Nationale Supérieure des Télécommunications
46 rue Barrault, 75634 Paris cedex 13, France
{Francois.Yvon, Gerard.Chollet}@enst.fr
[(3)] Jinny Software Ltd, Samra Center, 1st floor, Fanar, El Metn, Lebanon
Antoine.Ghaoui@jinny.ie

10.21437/Interspeech.2005-17

Abstract

State of the art Speech Recognition systems use statistical language modeling and in particular N-gram models to represent the language structure. The Arabic language has a rich morphology, which motivates the introduction of morphological constraints in the language model. Class-based N-gram models have shown satisfactory results, especially for language model adaptation and training from reduced datasets. They were also proven quite effective in their use of memory space. In this paper, we investigate a new morphological class-based language model. Morphological rules are used to derive the different words in a class from their stem. As morphological analyzer, a rule-based stemming method is proposed for the Arabic language. The language model has been evaluated on a database composed of articles from Lebanese newspaper Al-Nahar for the years 1998 and 1999. In addition, a linear interpolation between the N-gram model and the morphological model is also evaluated. Preliminary experiments detailed in this paper show satisfactory results.

1. Introduction

Language modeling aims at capturing local syntactic constraints between words. Statistical language models represent these constraints as probability distributions over words sequences. In the case of N-gram language models, the probability of a word sequence is computed as a product of the conditional probabilities of each word given a restricted history or context as follows:

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i \mid w_{i-1}, ..., w_{i-N+1}) \qquad (1)$$

For the i^{th} word, the context/history is limited to the previous N-1 words. Even with such restricted contexts, N-gram models include a huge number of parameters. A bi-gram (N=2) model on a vocabulary of V words has V^2 parameters $P(w_i|w_j)$, where $P(w_i|w_j)$ represents the probability to have word *j* followed by word *i*. The estimation and use of such a large parameter set raise several practical problems. First, the training corpus has to be large enough to estimate all these parameters. In fact, even if a very large corpus were available, not all the combinations would appear, since some of them are extremely rare. Secondly, N-gram models require a large memory space.

In order to overcome these problems, a parametric distribution function with a limited number of parameters is generally used for the conditional probability distribution of the vocabulary words. However, the use of a parametric distributions function requires a distance measure or at least an order to be defined on the set of vocabulary words. No such distance measure or order exists. A possible solution is to cluster the vocabulary words into classes and to consider that words in a given class share certain properties. The use of word classes smoothes the probability distribution of the (word based) N-gram model. Syntactical classes, semantic classes, morphological classes or classes induced directly from the data have been considered. In this work, we are particularly interested in morphological constraints as a basis for class-based N-gram modeling.

In our model, we consider a word to be uniquely derived from its root by application of a morphological rule. The probability of a word, given its context, thus combines two terms: one is the contribution of the root based N-gram model; the other is probability of application of the morphological rule. This model is introduced in Section 2.

Section 3 describes the morphological model where a transducer is used to generate the different rules. The rules are empirical. This morphological class-based model has been linearly combined to the classical N-gram model.

Our morphological class-based language model has been implemented within the SRILM toolkit [1]. Preliminary experiments have been conducted using the articles of newspaper Al-Nahar for the years 1998 and 1999 as database. The database, experiments and results are detailed in Section 4. The final section of the paper presents conclusions and perspectives.

2. Morphological class based Language Model

Class-based language models have been studied (see e.g. [2; 3; 4]). In this paper, a particular approach is proposed where morphological rules are used to define the classes. Morphologically rich languages have a large set of words that derive from a small set of roots. The basic idea in our morphological class based model is to define a class per root and to associate all derived words to that class. The underlying assumption is that a strong relationship exists between the roots of the words in a language. The meaning of a sentence is completed by adding to the roots the derivation rules information.

Let w_i be a word of the vocabulary, r_i the root of w_i and g_i the morphological rule which allows the word w_i to be derived from r_i. At this stage, we assume that a word derives from a unique root, which may not be true in some cases. Each possible root in the vocabulary defines a class, which contains all the words deriving from that root.

Consider an N-gram language model where, for a context formed from N-1 words, the corresponding conditional probability distribution of a vocabulary of words V is defined by the set:

$\{\Pr(w_n/w_{n-1}, w_{n-2}, \ldots, w_{n-N+1})\}$ where $w_i \in V$ (2)

Every word w_i is uniquely defined by its root r_i and the corresponding rule g_i, denoted by the pair $w_i = (r_i, g_i)$. The probability of a word given its context can be written as:

$$P(w_n / w_{n-1},...,w_{n-N+1}) = P\big((r_n,g_n)/(r_{n-1},g_{n-1}),...,(r_{n-N+1},g_{n-N+1})\big) = P(g_n / r_n, r_{n-1}, g_{n-1},..., r_{n-N+1}, g_{n-N+1}).\ P(r_n / r_{n-1}, g_{n-1},..., r_{n-N+1}, g_{n-N+1}) \quad (3)$$

Eq. 3 defines a class based N-gram model that is equivalent to the basic N-gram. It is just a rewriting of the N-gram model, taking the morphological decomposition of the vocabulary words into account. In what follows, successive approximations will be proposed, in order to permit a smoothing of the N-gram language model by integrating the morphological information. These approximations are mainly based on the rough association of the roots to the core semantics of a sentence and the association of the rules with the grammatical aspects and corresponding semantics.

The first approximation to be made (basic smoothing of the distribution) is that the probability of n^{th} root r_n only depends on the preceding N-1 roots ($r_{n-1}, \ldots, r_{n-N+1}$). This is an acceptable approximation since it states that, at a certain level, the dependency between meaning units is often uncorrelated with the grammatical relations. Eq. 3 thus becomes:

$$P(w_n / w_{n-1},...,w_{n-N+1}) \cong P(g_n / r_n, r_{n-1}, g_{n-1},..., r_{n-N+1}, g_{n-N+1}).\ P(r_n / r_{n-1},..., r_{n-N+1}) \quad (4)$$

(Eq. 4) defines a first morphological class-based language model. If the number of roots in the vocabulary V is n_r, and the number of possible rules is n_g (few hundreds), the total number of parameters in the model expressed by (4) is: $n_g n_r (n_r n_g)^{N-1} + n_r^N$. This model doesn't reduce the number of parameters compared to classical N-gram model.

A second approximation can be performed. The morphological rule for the word w_n may be considered to depend only on the type T(r) of the previous words' roots. With this approximation, the rules are associated to grammatical constraints. A new equation may be derived from Eq. 4:

$$P(w_n / w_{n-1},...,w_{n-N+1}) \cong P(g_n / r_n, T(r_{n-1}), g_{n-1},..., T(r_{n-N+1}), g_{n-N+1}).\ P(r_n / r_{n-1},..., r_{n-N+1}) \quad (5)$$

The number of parameters in this new model is equal to: $n_g n_r (n_g)^{N-1} + n_r^N = n_r n_g^N + n_r^N$.

Comparing this number to the number of parameters in a classical N-gram v^N and taking into consideration the following assumptions:

- $v = n_g n_r$,
- $n_r >> n_g$

We can conclude that the new model (Eq. 5) reduces the number of parameters by a factor of $O(n_g^N)$ wrt. the word based N-gram model.

This approximation may be extended if we accept that a rule depends only on the type of the root, e.g. whether it is a verb, noun, adverb, name etc. This yields:

$$P(w_n / w_{n-1},...,w_{n-N+1}) \cong P(g_n / T(r_n), T(r_{n-1}), g_{n-1},..., T(r_{n-N+1}), g_{n-N+1}).\ P(r_n / r_{n-1},..., r_{n-N+1}) \quad (6)$$

Finally, if we assume that the rule of word w_n is independent from the rules used to derive the preceding words $w_{n-1}, \ldots, w_1$ we get:

$$P(w_n / w_{n-1},...,w_{n-N+1}) \cong P(g_n / T(r_n)).P(r_n / r_{n-1},..., r_{n-N+1}) \quad (7)$$

(Eq. 6) is the model we have implemented and used in our experiments, where the rule of word w_n depends on the rules of the preceding words $w_{n-1}, \ldots, w_{n-N+1}$. This has been tested for a context length <= N the order of the original model. (Eq. 7) is a particular case where the context length is taken to be equal to 1.

3. Morphological Transducer

In the morphological class-based N-gram model we propose in this paper, the morphological analyzer is a basic module. Morphological analysis allows to determine the root or stem of a word and the corresponding rule permitting to derive this word from that stem. In this work, we have focused on the Arabic language because of its rich morphology. The morphologic analysis of the Arabic language has been largely studied in the past [5]. Several techniques based on finite-state transducers have been developed [6]. In the present work, we propose to use a simple automaton to represent elementary morphological rules. Words are generated from a given root using prefixes and suffixes that define the morphological transducer. This basic morphological transducer does not support affixes. Our transducer contains the following rules:

- Words that begin with AL (' ال', 'وال', 'فال', 'بال', 'كال', 'لل').
- Words that begin with AL and end with PLUR (' ان', 'ون', 'ين').

- Words that begin with GEN (‘ ,’ك‘ ,’ب‘ ,’ف‘ ,’و‘ ,’س‘ ,’ل‘) and end with POS (‘ ,’تموني‘ ,’تمونا‘ ,’تهما‘ ,’تموني‘ ,’تكَ‘ ,’تكِ‘ ,’تني‘ ,’تنا‘ ,’وهما‘ ,’تيني‘ ,’تينا‘ ,’تكن‘ ,’تهم‘ ,’تهن‘ ,’تها‘ ,’وكَ‘ ,’وكِ‘ ,’وكم‘ ,’وكن‘ ,’وها‘ ,’وهم‘ ,’وهن‘ ,’وني‘ ,’ونا‘ ,’هما‘ ,’تَ‘ ,’تِ‘ ,’ك‘ ,’ك‘ ,’كم‘ ,’كن‘ ,’ني‘ ,’ها‘ ,’هم‘ ,’هن‘ ,’وا‘ ,’وك‘ ,’وه‘ ,’تك‘ ,’ته‘ ,’تم‘ ,’نا‘ ,’تَ‘ ,’ك‘ ,’ه‘ ,’ي‘ ,’ت‘).
- Words that begin with GEN and end with PLUR.
- Words that end with POS.
- Words that end with PLUR.

To illustrate the model used, below is an example for the rule: AL+root

“الولد” (the child) = "ال+ولد" (the + child)

The implementation and the extraction of the root from the words are done using a graph transducer structure (‘‘Finite State Transducer’’ (FST)) as shown in Figure 1:

Root
AL_*root*
AL_*root*_PLUR
GEN_*root*
GEN_*root*_POS
GEN_*root*_PLUR
*root*_POS
*root*_PLUR

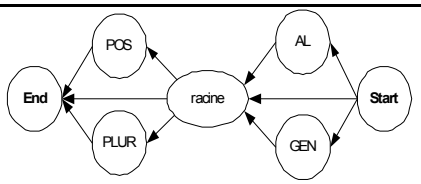

Figure 1: Transducer implementing morphological rules

4. Experiments and Results

The experiments are conducted using the SRILM toolkit [1] on a database composed of articles from newspaper Al-Nahar for years 1998 and 1999. SRILM software doesn’t integrate a solution to experiment the model proposed in this paper. Thus, an integration of the model within the software has been performed. The database and the proposed model implementation into SRILM is briefly described hereunder.

4.1. Database

The database used in the experiments is based on the articles of the Lebanese newspaper Al-Nahar for the years 1998 and 1999. In the present work, the database has been divided into two parts; year 1999 is used for the training stage and year 1998 for the testing. The articles are HTML based. A simple HTML parser has been developed to extract the text from the HTML pages. The training data, which correspond to the year 1999, contains 44234 HTML pages. The test data contains 47766 HTML pages. After pre-processing, the training part of the database contained 429229 different words with a frequency of occurrence varying from 1 to 754100. The test part is composed of 440298 different words where the frequency of appearance varies from 1 to 733023.

In our preliminary experiments, the vocabulary has been limited to those words with a frequency greater than 100. This produces a vocabulary of 18119 different words. Using the morphological analysis of section 3, the vocabulary words derive from 10269 different roots.

4.2. Implementation

SRILM [1] implements a solution for the class-based language model. The SRILM existing solution couldn’t directly be used to implement our model described in (6). Thus, the SRILM software has been modified to implement the model proposed in this paper (refer to Section 2).

The idea is:

- Estimate the N-gram on roots of the words using SRILM.
- Estimate the N-gram on rules from which the words are generated. The rules contain the type of the roots, which are limited to Verb and Non-Verb. Using the rules’ N-gram, Eq. 6 can be written:

$$P(w_n / w_{n-1},\ldots,w_{n-N+1}) \cong \frac{P(g_n, T(r_n)/g_{n-1},\ldots,g_{n-N+1},)}{\sum_{i/T(r_i)=T(r_n)} P(g_i, T(r_n)/g_{n-1},\ldots,g_{n-N+1})} \times \mathrm{P}(\mathrm{r_n}/\mathrm{r_{n-1}},\ldots,\mathrm{r_{n-N+1}}) \quad (8)$$

- Once the roots’ and the rules’ N-gram models are estimated, perplexity is computed on the test data. For this purpose, a modification is introduced to the software in order to implement (Eq. 8). For every word and context, the roots and rules are found. The N-gram probability of the root and the rules is determined using SRILM; these probability are then multiplied and divided by the summation, which is computed aside.

In order to smooth the N-gram model and to cope with unseen words/contexts pairs, back-off methods are generally used [7]. If a word is not found within a specific context (e.g. trigram), i.e. the couple word/context is not seen in the training set, it is checked against a larger context (e.g. bigram) until it is found. This process continues till the smallest context, the unigram. The probability from larger context has to be multiplied by a scaling factor, also called back-off factor. The back-off factor represents a small amount of the probability mass reserved for unseen words/contexts.

SRILM implements several back-off approaches. However, for the root N-gram the classical back-off methods are not applicable in the morphological class-based model. This is due to the fact that the probability is computed using (Eq. 8). The normalization factor is computed differently taking into account the probability distributions of the different rules.

4.3. Experiments results

Experiments have been conducted on the Al-Nahar database described in subsection 4.1. The trigam model has been applied with different context lengths for the rules. The performance of the model is measured using the perplexity. The perplexity measures an average branching factor for a given context. It is calculated as the exponential of the entropy of the N-gram distributions.

The classical trigram has a perplexity of 376.237 using the test data with an out-of-vocabulary rate of 17%. The trigram, using the morphological class-based with different context lengths for the rules’ N-gram, gives the perplexity in the table below:

Context length of the rules	Perplexity
1	1033.09
2	971.881
3	956.157

These results show that the morphological class-based model yields poor performance compared to a classical trigram. This is in line with results generally obtained with class-based N-gram models, and is due to the fact that the class-based model has a much smaller number of parameters. These results are expected, based on the different simplifications applied in our model of (Eq. 6).

4.4. Linear Interpolation

In order to investigate if the morphological class-based model may be combined to a classical N-gram model, linear interpolation has been used. The linear interpolation permits a reduction in the perplexity by a factor of 1.7% (369.893) as shown in the figure 2. This shows that our morphological class-based model can bring some improvement over the classical N-gram model.

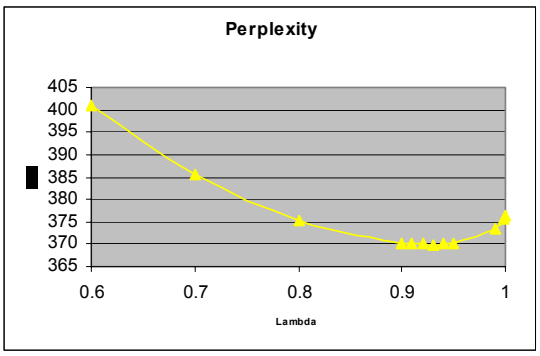

Figure 2: Results with linear interpolation between the morphological class-based model and the classical trigram model.

5. Conclusions and Perspective

In this paper an N-gram statistical language model that integrates morphological constraints in the form of classes is proposed. A complete theoretical framework is provided to integrate morphological constraints in the N-gram modeling. Depending on the assumptions made, multiple models may be derived from our theoretical framework.

A simplified model based on morphological classes has been applied to the Arabic language. A simple Arabic morphological analyzer has been developed for this purpose where empirical morphological rules are used in the form of a transducer. The SRILM toolkit has been modified to integrate the simplified model. Preliminary tests show a degradation of the performance in term of perplexity when compared to the classical N-gram. This was expected due to the reduction in the number of parameters in the modeling. The loss is term of perplexity is compensated by a reduction in the complexity of the model. The linear interpolation between the classical N-gram and the model based on morphological class improves the result in term of perplexity. This has been shown in our experiments.

This work will be pursued in various directions. First a more complete and accurate morphological analyzer will be used. Second, the morphological class-based language model will be experimented with fewer assumptions. Moreover, the assumption that a unique stem exists for a word will be relaxed by considering several stems per word with a probability associated to each of them.

6. Acknowledgements

This work is partially supported by the EC project NEMLAR.

7. References

[1] Stolcke A., "SRILM - An Extensible Language Modeling Toolkit,'' *Proc. Intl. Conf. Spoken Language Processing*, Denver, Colorado 2002.

[2] Manning C. and Schutze H, *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge 1999.

[3] Brown P. F., Della Pietra V. J., deSouza P. V., Lai J. C., and Mercer R. L., Class-based N-gram models of natural language, *Computational Linguistics*, Vol. 18, pp. 467–479 1992

[4] Neisler T, Category-based statistical language models, *Ph.D thesis, Department of Engineering, University of Cambridge*, U.K., 1997

[5] Darwish K, "Building a Shallow Arabic Morphological Analyzer in One Day" *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages,* Philadelphia, Boston, 2002.

[6] Beesley K, "Arabic Finite-State Morphological Analysis and Generation," *COLING-96*, Vol. 1, pp. 89-94 1996

[7] Katz S, "Estimation of Probabilities from Sparse Data for the Language Model Component of a speech Recognizer", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, no. 3, pp. 400-401 MARCH 1987