AN INFORMATION-BASED METHOD FOR SELECTING FEATURE TYPES FOR WORD PREDICTION

Dekai WU †, *SUI Zhifang* *†, *ZHAO Jun*†

† Human Language Technology Center,
Department of Computer Science,
University of Science & Technology, HKUST, Clear Water Bay, Hong Kong
* Computational Linguistics Institute,
Department of Computer Science & Technology,
Peking University, Beijing ,100871, P.R.China
{dekai,suizf,zhaojun}@cs.ust.hk

ABSTRACT

This paper uses an information-based approach to conduct feature types selection for language modeling in a systematic manner. We describe a quantitative analysis of the information gain and the information redundancy for various combinations of feature types inspired by both dependency structure and bigram structure through analyzing an English treebank corpus and taking word prediction as the object. The experiments yield several conclusions on the predictive value of several feature types and feature types combinations for word prediction, which are expected to provide reliable reference for feature type selection in language modeling.

1. INTRODUCTION

In the field of speech recognition, N-gram model is a practical model for its simplicity and efficiency. However, it still has some significant drawbacks, such as frequently linguistically implausible, and makes inefficient use of the training corpus[1][2]。On the other hand, variant structure-based language models have been developed[3][4]. However, the experiments show that the performance of the models based on pure structure information can not surpass or even match that of N-gram model. The goal of our research is to incorporate grammatically-based feature types into the N-gram model, which could incorporate the predictive power of words that lie outside of N-gram range without sacrificing the known performance advantages of N-gram models. Exactly, the ultimate measure of the quality of a language model is its impact on the application. However, discrepancies between training sets, evaluation criteria, algorithms, and hardware environments makes it difficult to compare the models objectively. In the paper, we use an information-based approach to conduct feature types selection in a systematic manner. We describe a quantitative analysis of the information gain and the information redundancy for various feature types combinations inspired by both dependency structure and bigram through analyzing Penn Treebank and taking word prediction as the object. The experiments yield several conclusions on the predictive value of several feature types and feature types combinations for word prediction, which are expected to provide reliable reference for feature type selection in language modeling.

2. INFORMATION-BASED MODEL FOR FEATURE TYPES SELECTION

A language model is used to predict a given word based on its history. By the laws of conditional probabilities, a language model can be represented in left-to-right fashion as

$$P(S) = P(w_0)P(w_1 \mid h_1)\cdots P(w_i \mid h_i)\cdots P(w_n \mid h_n)$$

where S denotes a sequence of words w_0, w_1, ... , w_n, and h_i denotes the history of w_i $(1 \leqslant i \leqslant n)$.

In order to construct a language model, the individual probabilities $p(w_i|h_i)$ should be estimated from the training set. Since there are too many possible histories but not enough evidence in the training set, several feature types must be used to divide the space of possible histories into equivalence classes via the map $\Phi : h_i \xrightarrow{F_1,F_2,.....,F_K} [h_i]$ to make the model feasible in the implementation. The candidate feature types can be physical position based ones, as in N-gram models, or grammatically-based ones, as in dependency structure. How to select the optimal

feature types combination from the various candidates is an important task. We use an information-based approach to conduct feature types selection for language modeling in a systematic manner. In the following, some related concepts are introduced from the viewpoint of information theory, which are adopted as the foundation for feature type analysis.

(1)Information gain (*IG*): The information gain of taking F_2 as a variant model on top of a baseline model employing F_1 for predicting word *O* is defined as the average mutual information[5] between the predicted word *O* and F_2, given that feature type F_1 is known.

$$IG(F_2;O \mid F_1) = E_{p(F_1F_2O)}\left[\log \frac{p(F_2O \mid F_1)}{p(F_2 \mid F_1)p(O \mid F_1)}\right]$$

(2)Information quantity (*IQ*): When the baseline model employs no feature type for word prediction, the information gain of taking *F* as a variant model can be referred as the information quantity of feature type *F* for predicting word *O*, which is the average mutual information between *F* and *O*.

$$IQ(F,O) = IG(F,O \mid null) = E_{FO}\left[\log \frac{p(FO)}{p(F)p(O)}\right]$$

(3) Information redundancy(*IR*): Based on the above two definitions, we can draw the definition of information redundancy. $IR(F_1,F_2;O)$ denotes the redundant information between F_1 and F_2 in predicting *O*, which is defined as the difference between $IQ(F_2;O)$ and $IG(F_2;O|F_1)$, or the difference between $IQ(F_1;O)$ and $IG(F_1;O|F_2)$.

$$\begin{aligned} &IR(F_1,F_2;O) \\ &= IQ(F_2;O) - IG(F_2;O \mid F_1) \\ &= IQ(F_1;O) - IG(F_1;O \mid F_2) \end{aligned}$$

In the following experiments, we will use information gain to select the feature types series, while information redundancy is used to analyze the overlap degree between the variant and the baseline.

3. IMPLEMENTATION

3.1 The Feature Types Employed in the Experiments

The feature types in our experiments of feature types selection come from dependency structure and bigram structure. As for structural feature types, we take dependency grammar as a framework, since it extends N-gram models more naturally than stochastic context-free grammars. The feature types we used are listed as Table 1.

Table 1: The feature types used in the training set

B	Nearest preceding word	BP	POS of B		
M	Nearest preceding word modifying O	MP	POS of M	MT	Modifying type between M and O
R	Nearest preceding word modified by O	RP	POS of R	RT	Modifying type between R and O

For example, in the sentence " It has no bearing on our work force today.", taking "bearing" as *O* (the predicted word), then *B* (the nearest preceding word of *O*) is "no", *M* (the nearest preceding word modifying *O*) is also "no", *R* (the nearest preceding word modified by *O*) is "has", *BP* is the POS of "no", that is *DT*, *MP* is the POS of *M* "no", that is *DT*, *RP* is the POS of *R* "has", that is *VBZ*, *MT* is the modifying type between *M* "no" and *O* "bearing", that is *NP*, *RT* is the modifying type between *R* "has" and *O* "bearing", that is *VP*.

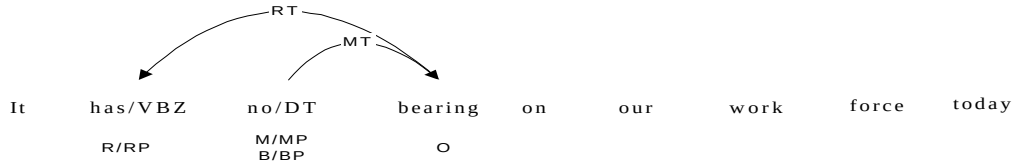

Figure 1: An example sentence to describe each feature type listed in Table 1

3.2 Experimental Method

The experiment is conducted on Penn TreeBank[6]. 80% of the corpus (979,767 words) is taken as the training set, which is used to estimate the probabilities used in computing information gain. 10% of the corpus (133,814) is taken as the testing set, which is used to compute the information gain.

The main obstacle in the procedure of computing

information gain is the zero-frequency problem. In our experiment, we use a blending approach to solve this problem, where the predictions of several contexts of different number of feature types are combined into a single overall probability.

Let L be the total number of candidate feature types, $F_1,F_2,...,F_i$ ($0<i<L$) be the feature type series selected till now, model-i be the language model which employs feature type series $F_1,F_2,...,F_i$ for word prediction, $p(F_1,F_2,...,F_i;O)$ be the probability assigned to O by model-i. If the weight given to the model-i is w_i, the blended probability $p(F_1,F_2,...,F_i;O)$ is computed by

$$p(F_1,F_2,...,F_i;O)$$
$$= w_{-1}p_{-1}(O) + w_0 p_0(O) + \sum_{j=1}^{i} w_j p(F_1,F_2,...,F_j;O)$$

where the weights should be normalized to sum to 1. $p_0(O)$ denotes the probability of O predicated by unigram model, $p_{-1}(O)$ denotes the probability of O predicated by the model in which every word is given an equal probability, that is the reciprocal of the vocabulary size.

We define the weights in the above equation according to the escape mechanism given in[7].

4. CONCLUSIONS

The experiments led to a number of interesting conclusions on the predictive power of various feature types and feature types combinations, some in support of traditional linguistic intuition and some more surprising. All of the conclusions can provide the heuristic information for language modeling.

4.1 Grammatically Motivated Feature Types Do Not Easily Yield as Much Predictive Information as Simple Bigrams.

From a traditional linguistics viewpoint, the feature types which describe structural in formation, such as *R* (the nearest preceding word modified by the predicted word *O*), *M* (the nearest preceding word modifying the predicted word *O*) etc., should be more significant for word prediction than the bigram predictor *B* (the nearest preceding word of the predicted word *O*). However, in the empirical information quantities shown in Table 2, the opposite turns out to be true. For B has the largest information gain in all of the feature types, bigram features outperform the grammatically-based features in the predictive power. The conclusion gives the explanation that none of the structural models have surpassed the n-gram approach, although the former is intuitively well-motivated.

4.2 Although *R* (the Word Modified by the Predicted Word) Is Less Effective Than *M* (the Word Modifying the Predicted Word) When They Are Used Individually for Word Prediction, *R* Is More Effective Than *M* If They Are Used on Top of a Standard Bigram Model (the Feature Type *B*).

To see this, inspect the following measurements from our experiments: $IQ(R;O)$=0.84603 bits which is less than $IQ(M;O)$=1.16932 bits, which shows that *M* is effective than R when used individually for word prediction. However, $IG(R;O|B)$=0.45755 bits which is greater than $IG(M;O|B)$=0.31048 bits. That is to say, taking *B* as the baseline, the prediction information for *O* brought by *R* is larger than that brought by *M*. Therefore, in principle, the language model which incorporates bigram and feature type *R* is likely to have the higher performance than the model which incorporates bigram and *M*.

4.3 If *M* (the Nearest Preceding Word Modifying the Predicted Word *O*) Is One of the Feature Types of the Baseline, *MT* (the Modifying Type between *M* and *O*) Will Bring Less Information Gain for Word Prediction.

We measured the information gain of *MT* over *M* to be only $IG(MT;O|M)$=0.11901 bits, while the information redundancy of *MT* and *M* is a much larger $IR(MT,M;O)$=0.69879 bits. This means that the prediction information for *O* in *M* ($IQ(M;O)$=1.16932 bits) contains almost all the prediction information for *O* in *MT*($IQ(MT;O)$=0.81780bits).

4.4 If *R* (the Nearest Preceding Word Modified by the Predicted Word *O*) Is One of the Feature Types of the Baseline, *RT* (the Modifying Type between *R* and *O*) Will Bring Less Information Gain for Word Prediction.

This simply mirrors the immediately preceding point (the claim in 4.3), except here *R* is the modified word (parent) instead of the modifying word (child). In this case, we measured the information gain of *RT* over *R* to be only $IG(RT;O|R)$=0.02169 bits, while the information redundancy of *RT* and *R* is a much larger $IR(RT,R;O)$=0.66161 bits. This means that the

prediction information for *O* in *R* (*IQ*(*R*;*O*)=0.84603bits) contains almost all the prediction information for *O* in *RT* (*IQ*(*RT*;*O*)=0.68330 bits).

4.5 Among the feature types in {*B, BP, M, MP, MT, R, RP, RT*}, the preference order for selecting feature types is *B, R, MT, BP, M, RT, RP, MP*.

We try to use the metric *IG* to obtain a feature type series for language modeling, which only considers performance gain while the complexity is ignored. To obtain this order, we performed a greedy search where at each step we selected the next most informative feature type (i.e., the feature type which has the largest information gain). The empirical information gain measurements in each searching step is shown in Table 2, where *IG*(*F*;*O*|*Null*)=*IQ*(*F*;*O*).

Table 2: Information gain measurements in a greedy search

baseline	Variant							
	B	*R*	*MT*	*BP*	*M*	*RT*	*RP*	*MP*
null	2.44021	0.84603	0.81780	1.57409	1.16932	0.68330	0.70982	0.91037
B	—	0.45755	0.36861	0.11321	0.31048	0.42851	0.42347	0.33483
B,R	—	—	0.23660	0.11207	0.20307	0.01776	0.01990	0.21451
B,R,MT	—	—	—	0.10655	0.01263	0.00743	0.00938	0.03075
B,R,MT,BP	—	—	—	—	0.00706	0.00552	0.00183	0.00461
B,R,MT,BP,M	—	—	—	—	—	0.00237	0.00047	0.00005
B,R,MT,BP,M,RT	—	—	—	—	—	—	0.00009	0.00004
B,R,MT,BP,M,RT,RP	—	—	—	—	—	—	—	0.00001

This preference ordering can serve as a reference for selecting feature type combinations in a language model. That is to say, given the feature type set {*B*, *BP*, *M*, *MP*, *MT*, *R*, *RP*, *RT*}, if a language model uses only one feature type, *B* should be it; if a language model uses two feature types, the feature type combination {*B*, *R*} should be used; and so on. However, we can see from table2 that the additional information gain falls off rapidly when more than three feature types are selected.

5. FUTURE WORKS

Based on the analysis, we will design, construct, and incrementally refine new language models for written and spoken English that incorporate varying levels of linguistic structure. These models will aim to capture regularities that arise from long-distance dependencies, which n-gram models cannot represent. At the same time, we will retain as many of the n-gram parameters as needed to capture important lexical dependencies.

6. REFERENCES

[1]A.Stolcke, C. Chelba, D.Engle, V.Jimenez, L.Mangu, H.Printz, E. Ristad, R. Rosenfeld, D.Wu, F.Jelinek and S. Khudanpur, "Dependency Language Modeling", 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report, Research Note 24, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, April 1997. http://www.speech.sri.com/people/stolcke/papers/ws96-report.ps.Z

[2]Ciprian Chelba, David Engle, Frederick Jelinek, Victor Jimenez, Sanjeev Khudanpur, Lidia Mangu, Harry Printz, Eric Ristad, Ronald Rosenfeld, Andreas Stolcke, Dekai Wu, "Structure And Performance of a Dependency Language Model", Proceeding of Eurospeech'97, 1997

[3]Michael John Collins, A New Statistical Based on Bigram Lexical Dependencies", In: Proceedings of the 34rd Annual Meeting of the Association for Computational Linguistics,1996.

[4]John Lafferty, Daniel Sleator, Davy Temperley, "Grammatical Trigrams: A Probabilistic Model of Link Grammar", Proceedings of the 1992 AAAI Fall

[5]Cover T. M., Thomas J. A., "Elements of Information Theory", John wiley and Sons Inc., New York, 1991.

[6]http://www.cis.upenn.edu/~treebank/home.html

[7]Bell, T.C., Cleary, J.G., Witten,I.H., Text Compression, PRENTICE HALL, Englewood Cliffs, New Jersey 07632, 1992