An Information-Theoretic Empirical Analysis of Dependency-Based Feature Types for Word Prediction Models

Dekai WU †, *ZHAO Jun*†, *SUI Zhifang* *†
† 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, zhaojun, suizf}@cs.ust.hk

Abstract

Over the years, many proposals have been made to incorporate assorted types of feature in language models. However, discrepancies between training sets, evaluation criteria, algorithms, and hardware environments make it difficult to compare the models objectively. In this paper, we take an information theoretic approach to select feature types 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, using a Chinese 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 guidelines for feature type selection in language modeling.

1 Introduction

There are many types of features that a language model can use to predict a word in a sentence. Standard n-gram models use the immediately preceding words. Other fixed physical distance feature types may inspect word classes or parts of speech. Grammatically-based feature types may also be used, such as the incident syntactic and semantic relations or the other words involved in those relations. Our ultimate aim is to determine which combination of feature types is optimal for language modeling. Unfortunately, the state of knowledge in this regard is very limited. Many language models have been published inspired by one or more of these feature types[1][2][3][4][5], but discrepancies between training sets, evaluation criteria, algorithms, and hardware environments make it difficult, if not impossible, to compare the models objectively. The paper uses an information theoretic approach to select feature types for language modeling in a systematic manner. We are concerned with quantitative analysis of the information quantity, information gain and the information redundancy for various feature type combinations in both dependency grammar structure and adjacent bigram structure. The experiments yield a number of conclusions on the predictive value of various feature types and the combinations thereof, which can provide useful information on what level of performance gain can be expected in principle from a bigram model augmented with long distance dependency features. The results are expected to provide a reliable reference for feature type selection in language modeling.

We have used Chinese data for the experiments in this paper. Strictly speaking, our conclusions apply only to Chinese. However, we actually expect very similar results on English, and all our preliminary experiments on English data do bear this out[6]. We believe the general methodology as well as many of the specific conclusions apply to a wide range of languages.

We will begin by introducing an information theoretic framework for feature type selection and analysis. We then describe the experimental setup. Finally, we discuss a number of claims deriving from the experimental evidence.

2 Framework

2.1 Features for Language Models

A language model predicts 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 $(0 < i \leq n)$.

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 $\Phi : h_i \xrightarrow{f_1, f_2, \ldots, f_K} [h_i]$ to make the model feasible in the implementation. In speech recognition, these feature types are most often fixed physical position based features, as in N-gram models. The feature types can be the words before the predicted word or the part-of-speech of the words before the predicted word. In order to remedy the linguistic implausibility and inefficient usage of the training set of N-gram models, we would like to incorporate grammatically-based feature types into the language model, which could incorporate the predictive power of words that lie outside of N-gram range[7][8]. However, we would like to do so without sacrificing the known performance advantages of N-gram models[9]. We follow the general approach of the aforementioned authors in taking dependency grammar as a framework, since it extends N-gram models more naturally than stochastic context-free grammars.

The feature types studied in this paper are combinations of the fixed physical distance features and grammatically based features listed in Table 1 and graphically depicted in Figure 1. To understand the feature types, consider the task of predicting "作业 (*zuo4 ye4*, assignment)" in the example sentence shown in Figure 2. We denote this word by O, which stands for "observed". The word bigram feature B is the nearest preceding word of O, in this case "英文 (*ying1 wen2*, English)". The nearest word modifying O is denoted by M, and is also "英文 (*ying1 wen2*, English)" in this case. Conversely, the nearest preceding word modified by O is denoted by R, "做 (*zuo4*, do)" here. BP is the part of speech of "英文 (*ying1 wen2*, English)", in this case "n(noun)". Similarly, MP is the POS of "英文(*ying1 wen2*, English)", and RP is the POS "v(verb)" for "做 (*zuo4*, do)". The modifying type or dependency relation between "英文 (*ying1 wen2*, English)" and "作业 (*zuo4 ye4*, assignment)" is denoted by MT, in this case "np(noun phrase)". RT is the modifying type between "做(*zuo4*, do)" and "作业(*zuo4 ye4*, assignment)", here "vp(verb phrase)".

Faced with so many feature types, one of the dilemmas for language modeling is which feature types, or feature type combinations, should be used. The experience has shown that the feature types should not be selected by intuition.

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

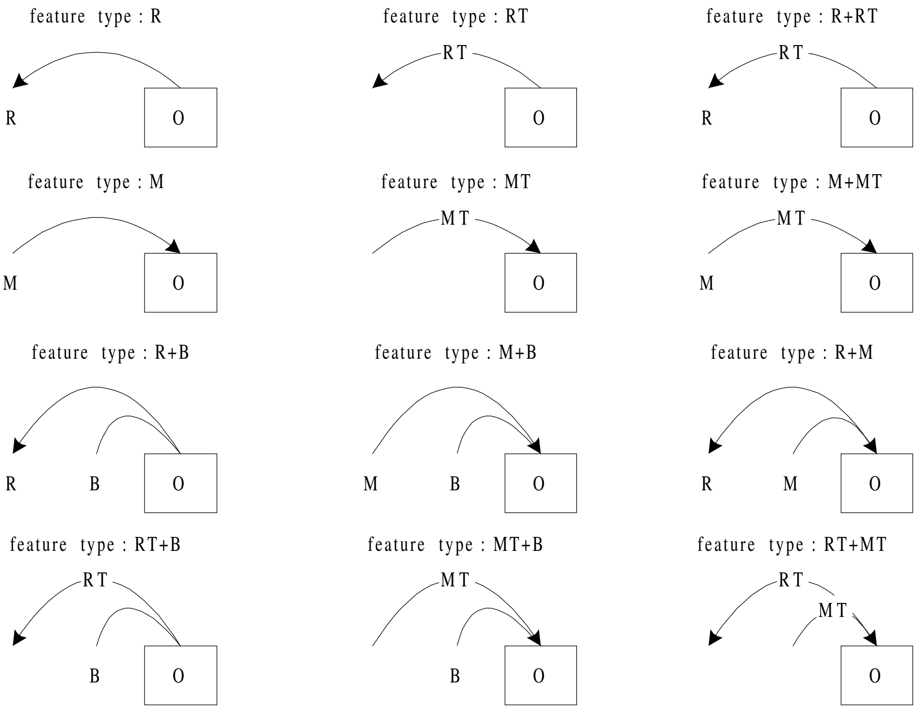

Figure-1: Some models using the combination of the bigram features and dependency-grrammar-based features

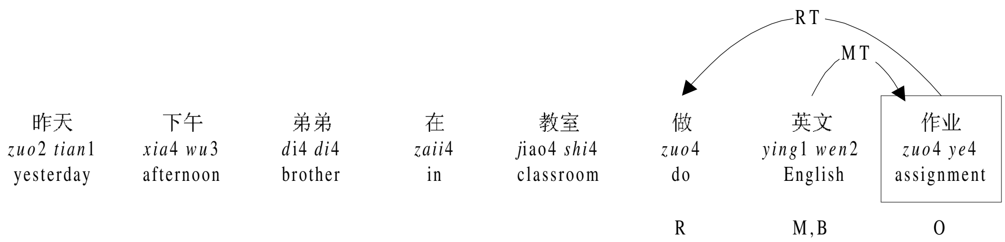

Figure 2: An example sentence to describe each feature type listed in Table 2

In order to obtain a more reliable reference to guide the addition of structural features to a stochastic language model, our objective is to establish in principle the amount of information available from various long distance dependency features and feature combinations. This can be regarded as an upper bound on the improvement that could be obtained by augmenting a language model with the corresponding features. We evaluate the informativeness of several feature types in bigram and dependency grammatical structure from the viewpoint of information theory. The experiments draw some conclusions on which feature types should be selected or should not be selected given specific baseline assumptions, and provide a ranking of the feature types according to their importance from this viewpoint.

2.2 Information-based Model for Feature Type Analysis

We now introduce some relevant concepts from information theory that we adopt as a foundation for analyzing feature types.

Information quantity (IQ). The information quantity of a feature type F to the predicted word O is defined using the standard definition of average mutual information[10]; we define IQ as the average mutual information between F and O.

$$IQ(F;O) = E_{p(FO)}\left[\log \frac{p(FO)}{p(F)p(O)}\right]$$

Information gain (IG). The information gain of adding F_2 on top of a baseline model that already employs F_1 for predicting word O is defined as the average mutual information between the predicted word O and feature type F_2, given that feature type F_1 is known.

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

Information redundancy (IR). The above two definitions lead naturally to a complementary concept 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 $IG(F_1;O|F_2)$.

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

We shall use IG to select the feature type series, and use IR to analyze the overlapped degree between the variant and the baseline.

3 The Corpus Used in the Experiments

The training corpus used in our experiments is a treebank consisting of Chinese primary school texts[11][12]. The basic statistics characterizing the training set are summarized in Table 2.

Table 2: Statistics of the training corpus

Corpus Size (words)	52,609
Total Sentences(sentences)	4,139
Average Sentence Length (words)	12.711
Vocabulary Size(words)	5,319
POS Tags	26
Phrase Types	14

In the experiments, we use 80% of the above corpus as a training set for estimating the various co-occurrence probabilities, while 10% of the corpus is used as a testing set to compute the information gain, information quantity, and information redundancy. The feature types we used in the experiments are those shown in Table 1.

4 Experimental Results and Analysis

Our experiments aim to quantitatively establish the amount of information intrinsically present in each feature type, and the information gain of each feature type on the top of various baselines. We were led to a number of conclusions on the predictive power of various feature types and feature types combinations, some in support of traditional linguistic intuition and some more surprising. These observations provide guidelines for language modeling. Below, we warm up with a well-known observation, and then move on to more focussed analysis.

4.1 Grammatically motivated feature types do not easily yield as much predictive information as simple bigrams.

From a traditional linguistics viewpoint, R (the nearest preceding word modified by the

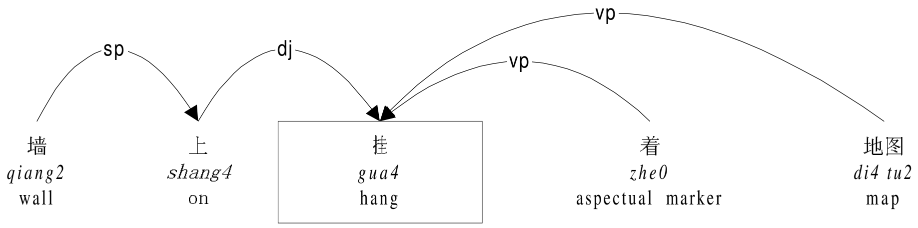

Figure 3: The dependency grammatical structure of Chinese sentence "墙/*qiang*2/wall 上/*shang*4/on 挂/*gua*4/hang 着/*zhe*0/(aspectual marker) 地图/*di*4 *tu*2/map." (There is a map hanging on the wall.)

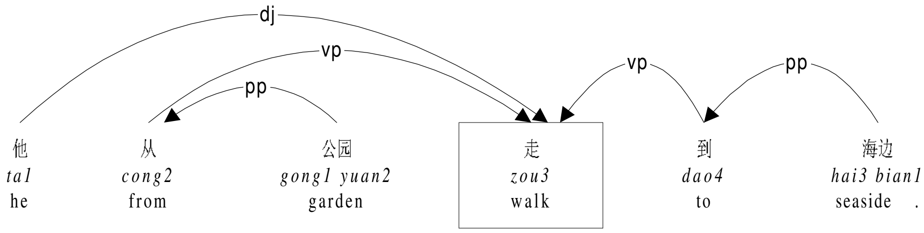

Figure 4: The dependency grammatical structure of Chinese sentence "他/*ta*1/he 从/*cong*2/from 公园/*gong*1*yuan*2/garden 走/*zou*3/walk 到/*dao*4/to 海边/*hai*3 *bian*1/seaside. (He walks from the garden to the seaside.)

predicted word O) should be more significant for word prediction than the bigram predictor B (the nearest preceding word of the predicted word O). Consider the sentence showed in Figure 3, where O is "地图/*di*4*tu*2/map", B is the aspectual marker "着/*zhe*0", and R is "挂/*gua*4/hang". It seems somehow obvious that R ("挂/*gua*4/hang") should be more predictive for O ("地图/*di*4*tu*2/map") than B (the aspectual marker "着/*zhe*0"). However, as is well known in speech recognition and statistical NLP research, the opposite turns out to be true. This is corroborated by the empirical information quantities shown in Table 3, which shows that B has the largest information quantity in all of the feature types. That bigram features outperform the grammatically-based features is commonly attributed to the predictive power of lexical association.

Table 3: Evidence for 4.1 (See text)

IQ(B;O)=3.826	IQ(MT;O)=0.971
IQ(M;O)=2.237	IQ(RT;O)=0.954
IQ(R;O)=1.581	IQ(MP;O)=0.818
IQ(BP;O)=1.493	IQ(RP;O)=0.663

Similarly, M (the nearest preceding word modifying the predicted word O) should be more significant for word prediction than B (the nearest preceding word of the predicted word O). For example, consider the sentence showed in Figure 4, where O is "走/*zou*3/walk", then B is "公园/*gong*1*yuan*2/garden" and M is "从/*cong*2/from". Again, it seems that M ("从/*cong*2/from") ought to be more predictive to O ("走/*zou*3/walk") than B ("公园/*gong*1*yuan*2/garden"), but from Table 3 we see that the opposite is true.

From a linguistic viewpoint, the explanation for the fact that R (IQ(R;O)=1.581) is less predictive than B (IQ(B;O)=3.826) may be as follows. Within a sentence, every word has

exactly one B and one R feature. But on one hand, the B feature always lies to the left of O since it is by definition the *preceding* word, while on the other hand, R generally lies to the right of O in Chinese sentences (with a few notable exceptions such as prepositional phrases). When R is not in the history preceding O, it cannot be used to predict O.
Similarly, a possible factor in the fact that M (IQ(M;O)=2.237) is less predictive than B is that M sometimes lies to the right of O. Another factor in the case of M is that none of the leaf nodes in a dependency tree have an M.

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 B).

Consider the following measurements from our experiments: IQ(R;O)=1.581 bits which is less than IQ(M;O)=2.237 bits, whereas IG(R;O|B)=0.683 bits which is greater than IG(M;O|B)=0.541 bits. That is, given a baseline bigram model employing only B features, augmenting the model with R features brings more information than augmenting it with M features. Therefore, in principle, the language model which incorporates bigram and feature type R can achieve higher performance than the model which incorporates bigram and M.

We believe this because there is more information redundancy between M and B than between R and B. From the above data, we see that there exists large information redundancy both between B and R (IR(B,R;O)=0.898) and between B and M (IR(B,M;O)=1.696). One explanation is that often B and M are in fact the same word, where the nearest preceding word modifies the predicted word. For example, consider the sentence in Figure 5, where "作业/*zuo*4*ye*4/assignment" is the predicted O, and B and M are the same word "英文/*ying*4*wen*2/English".

It is also possible that B and R are the same word, where the nearest preceding word is modified by the predicted word. For example, the dependency grammatical structure of the phrase "在/*zai*4/in 教室/*jiao*4*shi*4/classroom" is showed in Figure 6. Here, "教室/*jiao*4*shi*4/classroom" is the predicted O, and B and R are the same word "在/*zai*4/in".

In Chinese (as well as in English), the head word typically lies at the end of the phrase. This makes B more likely to be M than R, so the information redundancy between B and M is larger than that between B and R.

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 are interested in knowing how much non-redundant information is present in MT if M is included in the baseline. To assess this, we conducted the following experiment, which focuses directly on the relationship between MT and the two words involved.

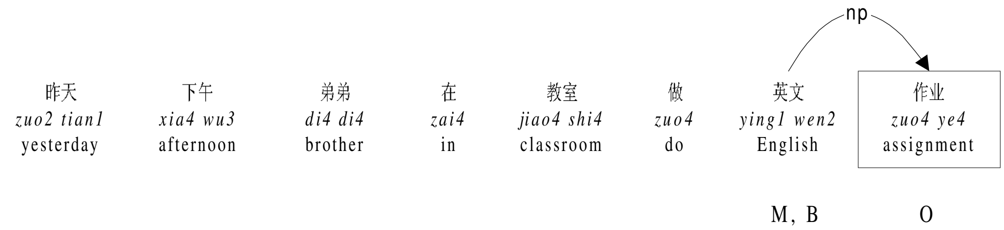

Figure 5: The dependency grammatical structure of "英文/*ying*1*wen*2/English 作业/*zuo*4*ye*4/assignment "

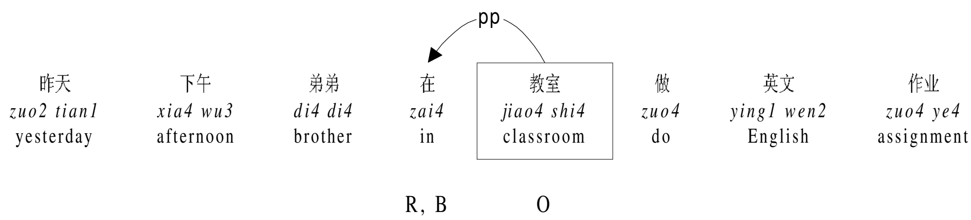

Figure 6: The dependency structure of "在/*zai*4/in 教室/*jiao*4*shi*4/classroom"

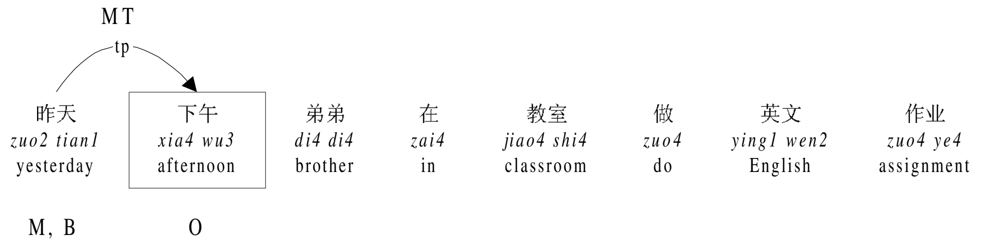

Figure 7: The dependency structure of the phrase "昨天/zuo2tian1/yesterday 下午/xia4wu3/afternoon"

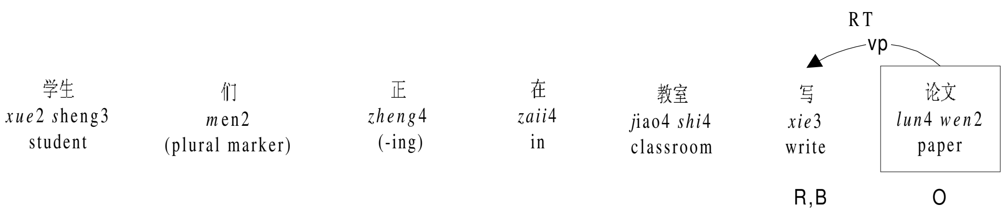

Figure 8: The dependency structure of the phrase"写/xie3/write 论文/lun4wen2/paper"

We measured the information gain of MT over M to be only IG(MT;O|M)=0.110 bits, while the information redundancy of MT and M is a much larger IR(MT,M;O)=0.861 bits. This means that the prediction information for O in M (which at IQ(M;O)=2.237 bits is much larger, incidentally, than that in MT at IQ(MT;O)=0.971 bits) contains almost all the prediction information for O in MT. The corresponding

Table4: Information gain measurements in a greedy search

Baseline	Information Gain of the Variants							
	B	R	M	RT	MT	BP	RP	MP
null	**3.826**	1.581	2.237	0.954	0.971	1.493	0.663	0.818
B	—	**0.683**	0.541	0.585	0.533	0.108	0.499	0.473
B,R	—	—	**0.388**	0.093	0.381	0.089	0.033	0.331
B,R,M	—	—	—	**0.084**	0.061	0.063	0.031	0.014
B,R,M,RT	—	—	—	—	**0.059**	0.052	0.009	0.011
B,R,M,RT,MT	—	—	—	—	—	**0.046**	0.008	0.007
B,R,M,RT,MT,BP	—	—	—	—	—	—	**0.007**	0.003
B,R,M,RT,MT,BP,RP	—	—	—	—	—	—	—	**0.002**

linguistic explanation may be as follows. The lexical identities of the predicted word O and its modifying word M involved in a dependency relation determine to a large extent the type of modification relation MT that holds between O and its modifying word M.

Consider the sentence in Figure 7. In the phrase "昨天/zuo2tian1/yesterday 下午/xia4wu3/afternoon", just knowing the identity of the two words "昨天/zuo2tian1/yesterday" and "下午/xia4wu3/afternoon" is enough to predict with near certainty that the relation between them is time phrase (tp), thus giving the following dependency structure as Figure 7.

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, except that 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)=0.271 bits, while the information redundancy of RT and R is a much larger IR(RT,R;O)=0.683 bits. This means that the information in R (IQ(R;O)=1.581 bits) contains almost all the information in MT (IQ(RT;O)=0.954 bits). The corresponding linguistic explanation is as follows. The lexical identities of the words (R, O) involved in a dependency relation determine to a large extent the type of modification relation RT that holds between O and the word it modifies, R.

Consider the sentence in Figure 8, the identity of the words "写/xie3/write" and "论文/lun4wen2/paper" determine with near certainty that their relationship is verb phrase (vp):

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, M, RT, MT, BP, RP, MP.

We used the metric IG to obtain a ranking for feature types according to their predictiveness. This ranking only considers information gain; it ignores complexity (for a practical application, we would also consider the complexity of the model at the same time.). 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 that has the largest information gain). The empirical information gain measurements in each search step is shown in Table 4, where the feature which has the boldface IG in each column is the feature type selected in that step, and IG(F;O|Null)=IQ(F;O).

This preference ordering can serve as a guideline for selecting feature type combinations in a language model. That is to say, given the

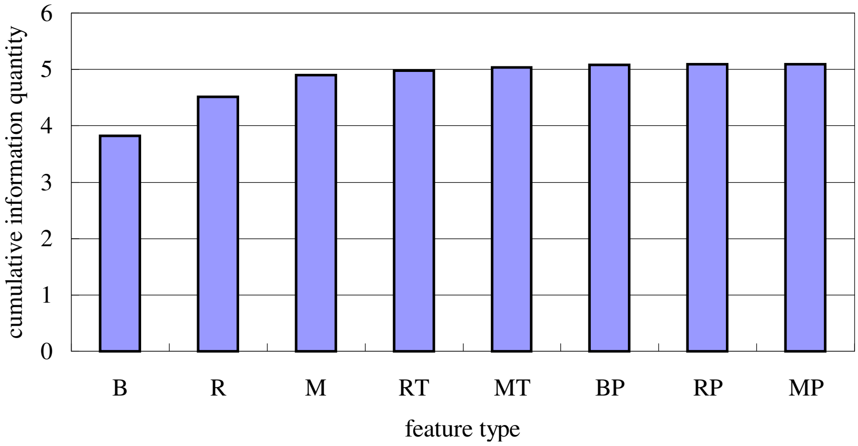

Figure 9: Cumulative information quantity of selected feature type combinations with 1~8 feature types

feature type set {B, BP, M, MP, MT, R, RP, RT}, if a language model uses only one feature type, feature type B should be used; 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 Figure 9 that the additional information gain falls off rapidly when more than three feature types are selected.

5. Conclusion

We have described a series of corpus-based analyses that take a Chinese treebank and quantify the information gain and the information redundancy for various feature types combinations involving both dependency and bigram feature types. The analysis yields several interesting conclusions that explain linguistic observations from an information theoretic point of view, and in addition will find practical use in the design of language models. Although perhaps we have been aware of some of the observations to varying extents, here we introduce a methodology that uses concrete evidence drawn from real contexts in order to give more reliable and objective results.

We have already begun conducting similar experiments on an English training corpus[6], which so far yield the same types of behavior described in this paper. We aim to discover which, if any, claims about the information present in dependency based features are peculiar to Chinese language, which are peculiar to English, and which are common across multiple languages.

Based on the analysis, we will design, construct, and incrementally refine new language models for written and spoken English and Chinese 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.

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.

[2] Della Pietra, S. and V. Della Pietra, "Inducing features of random fields", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(4), April 1997, pp.380-393.

[3] Ciprian Chelba, David Engle, Frederick Jelinek, Victor Jimenex, Sanjeev Khudanpur, Lidia Mangu, Harry Printz, Eric Ristad, Ronald Rosenfeld, Andreas Stolcke, Dekai Wu, "Structure and performance of a dependency language model", *Proceedings of Eurospeech'97*, 1997.

[4] Ney, Hermann., "On structuring probabilistic dependency in stochastic language modeling", *Computer Speech & Language* **8**: 1-38, 1994.

[5] John Lafferty, Daniel Sleator, Davy Temperley, "Grammatical trigrams: A probabilistic model of link grammar", *Proceedings of the 1992 AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, Cambridge, MA, 1992.

[6] Dekai WU, SUI Zhifang, ZHAO Jun, "An information-based method for selecting feature types for word prediction", to appear in *Proceeding of Eurospeech'99*, 1999.

[7] Michael John Collins, "A new statistical parser based on bigram lexical dependencies", in: *Proceedings of the 34rd Annual Meeting of the Association for Computational Linguistics*, 1996.

[8] Michael Collins, "Three generative, lexicalised models for statistical parsing", in: *Proceedings of the 35rd Annual Meeting of the Association for Computational Linguistics*, 1997.

[9] S. Della Pietra, V. Della Pietra, J. Gillett, J. Lafferty, H. Printz, L. Ures, "Inference and estimation of a long-range trigram model", 1994.

[10] Cover T. M., Thomas J. A., *Elements of Information Theory*, Wiley., New York, 1991.

[11] ZHOU Qiang, *Phrase Bracketing and Annotating on Chinese Language Corpus*, Dissertation for Doctor Degree [Peking University], Beijing, China, 1996.

[12] YU Shiwen, ZHOU Qiang, ZHANG Wei, ZHANG Yunyun, ZHAN Weidong, CHANG Baobao, SUI Zhifang, "Tagged Singapore Chinese primary school text", *Communications of COLIPS* **5**, 1995.