Using Conditional Random Fields to Predict Focus Word Pair in Spontaneous Spoken English

Xiao Zang[1,2], *Zhiyong Wu*[1,2,3], *Helen Meng*[1,3], *Jia Jia*[1,2], *Lianhong Cai*[1,2]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen Key Laboratory of Information Science and Technology,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2] Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

zangxiaocs@163.com, {zywu,hmmeng}@se.cuhk.edu.hk, {jjia,clh-dcs}@tsinghua.edu.cn

Abstract

This paper addresses the problem of automatically labeling focus word pairs in spontaneous spoken English, where a focus word pair refers to salient part of text or speech and the word motivating it. The prediction of focus word pairs is important for speech applications such as expressive text-to-speech (TTS) synthesis and speech recognition. It can also help in better textual and intention understanding for spoken dialog systems. Traditional approaches such as support vector machines (SVMs) prediction neglect the dependency between words and meet the obstacle of the imbalanced distribution of positive and negative samples of dataset. This paper introduces conditional random fields (CRFs) to the task of automatically predicting focus word pair from lexical, syntactic and semantic features. Furthermore, several new features related to syntactic and semantic information are proposed to achieve better performance. Experiments on the publicly available Switchboard corpus demonstrate that CRF model outperforms the baseline and SVM model for focus word pair prediction, and newly proposed features can further improve performance for CRF based predictor. Specifically, compared to the low recall rate of 11.31% achieved by the SVM model, the proposed CRF based predictor can yield a high recall rate of 70.88% with little impact on precision.

Index Terms: focus word pair, focus prediction, conditional random fields (CRFs), support vector machines (SVMs)

1. Introduction

The salience of information in spontaneous spoken English is conveyed by focus. The prediction of focus using text-based features in an utterance can help find words that should stand out with respect to their surrounding words. These focus words should be emphasized in expressive speech synthesis to accurately draw the attention of the listener. The location of the focus words is also important for spontaneous spoken language understanding systems [1]. Focus words can convey the intention of the speaker. The prediction of focus word pair can help spoken dialog systems to better understand the intention of the user. There is no standard definition for focus. Intuitively focus is the salient part of the text or speech. In this paper, we define focus as implying a set of alternatives to the focused word in the context by following the widely accepted alternative semantics definition [2]. In [3], the focus under this definition is called *kontrast* and is defined from a semantics perspective as six categories: contrastive, subset, correction, adverbial, answer and other. For example, (1) and (2) are samples of contrastive and subset kontrasts respectively.

- I have some in the **backyard** but I like those in the **front**.
- This woman owns **three day cares**, but she has to open **the second one** up.

There are only few researches on automatic prediction of focus in the literature. In [4], the regression classifier is used with the syntactic features, accentual features and word level acoustic features to predict focus. However, the main purpose of [4] is to explore the relationship between information structure and prosodic structure, but not to pursue a high accuracy of focus prediction. Our objective in this paper is to investigate the methods and features to predict focus from input text, which will be crucial for expressive TTS. Furthermore, the features used in [4] are manually labeled and cannot be automatically extracted from the text.

One major difficulty of predicting focus from text is that it is hard to extract rich enough features from a single word. Considering a focus word generally has one nearby word connected with it, we seek to predict the focus based on the context of focus word pair. The focus word pair is composed of the focus word and the word motivating it. It is easy to adopt the predicted focus word pair in the application scenarios of focus.

Contrastive word pairs, as part of focus word pairs, has been studied. Part-of-speech (POS), semantic similarity and acoustic features such as F0, duration, energy, etc. are used in the framework of decision tree to predict symmetric contrast [1], which is similar to the contrastive kontrast. Informative features including lexical, syntactic and semantic features are extracted from the text and used in support vector machine (SVM) to predict the contrastive word pair [5].

Following [5], we firstly attempt to adopt SVM to predict focus word pair from lexical, syntactic and semantic features. To deal with the imbalanced distribution between the positive and negative samples, the synthetic minority over-sampling technique (SMOTE) [6] is used to over-sample the positive instance class. However, a bad performance, especially a poor recall rate, is achieved. The reason might be due to that only the dependency between the focus word and the word motivating is considered in the SVM based predictor. Richer contextual information such as the dependency between the other words in the sentence and the focus word is neglected. Meanwhile, there is no good enough method to help SVM model solve the problem of imbalanced data distribution.

This paper introduces conditional random fields (CRFs) for the task of automatically predicting focus word pair. As a kind of sequence model, CRFs could take advantage of richer contextual information and predict a label for a single word with regard to neighboring words. CRFs are also less affected by the imbalanced data distribution. The rest of this paper is organized as follows. Section 2 introduces the corpus and the approach of data pre-processing. The features and CRFs are detailed in Section 3 and Section 4. The experimental setup and results are then reported in Section 5. Finally, Section 6 presents discussions and conclusions.

2. Corpus and data preparation

A subset of Switchboard corpus [7] is used in our experiment. Part of the corpus has been annotated with the information of focus [3], which can be used as the label. It has also been annotated with the syntactic structure [8] which is used in the process of data pre-processing. In the corpus, the focus is annotated in the form of kontrast marking and trigger. The word annotated as kontrast is the focus. Besides, the word motivating the kontrast marking is also annotated. The link between these two words is called trigger. The kontrast is annotated in two levels: word and noun phrase. We only consider the word level kontrast in our experiment. The sentences containing at least one kontrast marking are selected in our experiment.

The label we are predicting is a binary distinction of having focus or not. For each sentence selected from the Switchboard corpus, the word pairs are constructed for each word. If one word is annotated as kontrast, the word linked with it by the trigger is selected to form the word pair with it. This word pair is considered as the positive instance, namely *the focus word pair*. If a word is not annotated as kontrast, the first word sharing the same broad POS is selected to form the word pair with it. Here the "first" word is counted from the current word till the end of the sentence and then resume from the beginning of the sentence. This word pair is considered as the negative instance. If one word has no other word sharing the same broad POS with it in the sentence, it will be neglected. Stop words such as "*the*", "*but*" and "*and*" are removed before processing. Table 1 shows the example value of the process. The words in trigger (*find-look*) form the focus word pair and is given value +1. Every other word and the first word sharing the same broad POS with it form the negative instance and is given value -1.

A focus word pair might be contrastive word pair, subset word pair, correction word pair, answer word pair, adverbial word pair, etc. and all these are annotated in the corpus [7]. Only the subset, correction and contrastive word pairs are included in our experiment due to the restriction of same POS in word pair. Among these types, subset word pair and contrastive word pair are most frequent, as shown in Table 2.

Table 1. *Example values generated from the sentence: "You/PRP can/MD find/v a/DT lot/n of/IN good/j public/j schools/n if/IN you/PRP look/v real/r hard/r".*

Word1	Word2	Value
good	public	-1
real	hard	-1
find	look	+1
…	…	…
of	if	-1

Table 2. *Statistics of the types of focus word pairs in the corpus.*

Type	Number	Frequent
Subset	1363	33.2%
Contrastive	1325	32.2%
Adverbial	1304	31.8%
Correction	105	2.6%
Other	4	0.09%
Answer	2	0.05%

3. Features

3.1. Common features

The features considered for the prediction of focus word pair are all text-based. All the features used for the detection of contrastive word pair in [5] are included in our experiments. These features could be grouped into three main categories: lexical features, syntactic features and semantic features.

To simplify our description, the two words of word pair are referred to as W_1 and W_2, where W_1 precedes W_2 in the sentence. Examples of the common features are shown below:

Lexical features:

- Conjunctions, adverbs and prepositions (CAP words), e.g. "rather than" and "or", that activate a focus word pair.
- Degree of textual parallelism [9] between two sub sentences.

Syntactic features:

- Part-of-speech (POS) information such as if W_1 is the only word having the same broad POS as W_2 in the sentence.
- Dependency syntax such as if W_1 and W_2 have the same type of dependency relation (subject of, object of …).

Semantic features:

- The WordNet [10] semantic relation between W_1 and W_2, such as hypernyms (*chair-furniture*), antonyms (*good-bad*), entails (*show-see*), etc.
- The semantic similarity [10] between W_1 and W_2.

3.2. Newly proposed features

Besides, the peculiarity of subset word pair, which has the most instances in the focus word pair, should not be ignored. Some new features should be added in the features set to consider the characteristics of the subset word pair.

In subset word pair, W_1 is a member of W_2. CAP words like "rather than" or "or" may not activate subset word pair. However, sentence patterns like "a W_1 of W_2" or "W1 is a W2" often activate a subset word pair. So the determiner and copula preceding or connecting W_1 and W_2 should be considered as features. Attribute clauses also activate subset word pairs. If a sentence pattern like "W_1 is W_2 that …" occurs, W_1 and W_2 are more likely to form a subset word pair. Therefore, an attribute relation should also be a feature.

We find some words themselves may activate the relation of subset. For example, the occurring of word "somebody" implies a general set, and it also suggests that there may be a word which can form subset relation with "somebody" in the sentence. Statistics of the appearance number of the words occurred in a focus word pair in the dataset is shown in Table 3. As can be seen, some words like "here" and "he" occur to be part of focus word pair many times. So the word identity can be one useful feature. We map W_1 and W_2 to numbers with Hash function, and use these numbers as new features.

As the syntactic patterns of subset word pairs are different from that of contrastive, we need to extract some new features from syntactic dependency to consider this difference. It seems to be helpful to identify the subset word pair knowing that W_1 is head of W_2 or that the type of dependency of W_2. For instance, "example" and "invasion" form a subset word pair in the sentence "an example of invasion". "Example" is the head of "invasion". This information and the type of dependency are helpful to identify the subset relation between "example" and "invasion". These should be added into the features extracted from syntactic dependency as new features.

Table 3. *Appearance number of the words occurred in a focus word pair in the dataset.*

Word	Number	Word	Number
here	15	men	13
he	15	one	13
she	13	doing	11
somebody	13	question	11
people	13	reason	11

3.3. Summary of features

To sum up, all the text-based features used for automatic prediction of focus word pair are listed below:

- **Non-functional words:** the first CAP word preceding W_1 or W_2 and its distance; the first two CAP words preceding W_1 or W_2 and its distance; the CAP words between W_1 and W_2; the determiner and copula preceding W_1 or W_2 and its distance; the determiner and copula between W_1 and W_2. Here the distance means the number of words between the two words.
- **Conjugate:** whether W_1 is the conjugate of W_2.
- **Information of sub sentence:** whether W_1 and W_2 are in the same sub sentence; the Wagner & Fischer edit distance [9] between the sub sentences containing W1 and W2.
- **Part-of-speech:** the POS of W_1 and W_2; whether W_1 is the only word which has the same POS with W_2 in the sentence; whether W_1 is nearest with W_2 in the set of words which has the same POS with W_2.
- **Syntactic dependency:** whether W_1 is the head of W_2; the type of syntactic dependency of W_1 and W_2; whether the heads of W_1 and W_2 refer to the same item.
- **Semantic relation:** the WordNet relation between W_1 and W_2; the semantic similarity between W_1 and W_2. Here the semantic similarity is calculated by [10].
- **Word:** the hash value of W_1 and W_2; the word itself which word pair originates from.

4. CRFs for focus word pair prediction

A conditional random field (CRF) can be considered as a structured output extension of logistic regression. The model used in this paper is actually a linear chain CRF. It provides an efficient framework for sequence labeling. A linear chain CRF defines a conditional probability distribution of a label sequence ***y*** given an observation sequence ***x***, which takes the parameter form:

$$p(y\,|\,x) = \frac{1}{Z(x)} \exp\left(\sum_i \left[\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i)\right]\right) \quad (1)$$

where $t_j(y_{i\text{-}1},y_i,x,i)$ is a transition feature function of the entire observation sequence and the labels at position i and i-1 in the label sequence, and $s_k(y_i,x,i)$ is a state feature function of the label at position i and the observation sequence; λ_j and μ_k are the corresponding weights of these two functions and also the parameters to be estimated from the training data; $Z(x)$ is a normalization factor and is computed as:

$$Z(x) = \sum_y \exp\left(\sum_{i,j} \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_{i,k} \mu_k s_k(y_i, x, i)\right) \quad (2)$$

An ordinary classifier such as SVM predicts a label for a single word pair without considering neighboring word pairs, which is the major restriction for performance improvement. Linear chain CRF can overcome the restriction by taking into account the context information for the problem of sequence labeling. Due to the reason, CRF has been introduced for the task of focus prediction in our work.

The CRF++ tool [11] has been adopted to predict focus as it allows us to redefine feature sets and specify the feature templates in a flexible manner. In our work, feature templates are defined so that unigram, bigram and trigram contextual features can be used by the CRF predictor. These unigram, bigram and trigram contextual features can help CRF take richer context into account. Some of the templates are shown in Table 4. For example, a series of feature candidates are generated for each word after the process of feature extraction above. The feature template $X_{\text{-}1k}X_{0k}$ will generate a new bigram feature for each word pair that is composed of the k-th feature candidate of the previous word pair and the k-th feature candidate of the current word pair.

Table 4. *Definition and examples of features templates.*

Type	Templates	Definition
Unigram	$X_{\text{-}1k}$, X_{0k}, X_{1k}	The k-th feature candidates of previous [-1], current [0] and next [1] word pair
Bigram	$X_{\text{-}1k}X_{0k}$, $X_{0k}X_{1k}$	The combination of the k-th feature candidates of previous [-1] (next [1]) and current [0] word pair
Trigram	$X_{\text{-}1k}X_{0k}X_{1k}$	The combination of the k-th feature candidates of previous [-1], current [0] and next [1] word pair

5. Experiments and results

Two objective experiments are conducted to evaluate the performance of the proposed CRF based tagger for automatic predication of focus word pair. The first experiment is the comparison between the SVM and CRF using the proposed features. The second experiment validate the effectiveness of the newly proposed features by comparing the performance of different features using the proposed CRF method.

Accuracy, precision and recall are used as the performance measurement. They are defined as:

$$accuracy = (TP + TN)/(P + N) \quad (3)$$

$$precision = TP/(TP + FP) \quad (4)$$

$$recall = TP/(TP + FN) \quad (5)$$

where P is the number of all the positive instances in the test set; N is the number of all the negative instances in the test set;

TP is the number of positive instances which are correctly tagged as positive; *FP* is the number of negative instances which are incorrectly tagged as positive; *FN* is the number of positive instances which are incorrectly tagged as negative.

In the experiments, 70,767 instances are used, in which 1,583 instances are positive. 5-fold cross-validation is adopted for the experiments with different models or features. The data set is randomly and equally divided into 5 parts, from which one part is selected as the test set and the other four parts are used as the training set. The selection of the test set is repeated until all the data have been covered in the test set.

In the first experiment, LIBLINEAR implementation [12] is adopted for SVM based predictor. The synthetic minority over-sampling technique (SMOTE) [6] is used to solve the problem of imbalanced data distribution between positive and negative samples for SVM. To compare the performance between CRF and SVM with or without new features, the experiments are conducted on two feature sets. One feature set includes all the features mentioned in [5], which is called as *old feature set*, and the other set contains the new features proposed in this paper in addition to the old features, which is called as *new feature set*. Table 5 shows the results of the comparison experiments, and the experiments of SVM include one dealt with SMOTE ($\text{SVM}_{\text{SMOTE}}$) and one without (SVM). In SMOTE, the number of nearest neighbors is set to be 15 and the number of the synthesized positive instances is set to be the same as the positive instances in the original training set. The baseline is the tagger which always predicts the instances as negative. As can be seen, the recalls of SVM are very low on both the old and new feature set. For the new feature set, the recalls are 11.31% and 6.44% for SVM with and without SMOTE, while CRF performs significantly better at 70.88%. Using the method of SMOTE can increase the recall of SVM from 6.44% to 11.31%, but it has a bad impact to the precision which declines from 90.27% to 58.69%. But the CRF model can significantly improve the recall and has a very little impact to the precision. This trend is also the same on the old feature set. The results demonstrate that CRF based predictor outperforms SVM based predictor on both feature sets and the CRF based predictor can obtain a significantly better recall rate.

Table 5. *Performance comparison between SVM based predicator with/without SMOTE and CRF based predicator on old and new feature sets; the baseline is a predictor that always labels instances as negative.*

Model	Feature Set	Accuracy	Precision	Recall
Baseline	---	88.86%	---	---
SVM	old	89.43%	**87.85%**	5.93%
$\text{SVM}_{\text{SMOTE}}$	old	89.20%	58.49%	10.11%
CRF	old	**93.60%**	73.80%	**66.01%**
SVM	new	89.50%	**90.27%**	6.44%
$\text{SVM}_{\text{SMOTE}}$	new	89.24%	58.69%	11.31%
CRF	new	**94.98%**	81.66%	**70.88%**

Table 6 shows the experimental results of CRF based tagger with different sets of features. To simplify our description, we define whether W_1 is the head of W_2, and the type of syntactic dependency of W_1 and W_2 as *new dependency features*; define the determiner and copula preceding W_1 or W_2 and its distance, and the determiner and copula between W_1 and W_2 as *new non-functional features*; and define the hash value of W1 and W2 as *new word features*. From the results, it is obvious that the performance for the accuracy, precision and recall keeps improving every time a new feature is added. Specifically, accuracy increases 4.74% as compared to the baseline when all the features except new dependency features, new non-functional features, and new word features are used; it increases by another 0.22% after new dependency features are added; it further improves 0.62% when new non-functional features are added; and it finally reaches 94.98% when all the features are considered. Precision and recall have similar tendency. The experimental results suggest that the new dependency features, the new non-functional features, and the new word features can improve the performance. Such results can be attributed to that the performance of predicting the subset word pair is improved by using these newly proposed features.

Table 6. *Performance of focus word pair predicator based on CRF using different features; the baseline is a predicator that always labels instances as negative.*

Features	Accuracy	Precision	Recall
Baseline	88.86%	---	---
All the features – (new dependency features, new non-functional features, new word features)	93.60%	73.80%	66.01%
All the features – (new non-functional features, new word features)	93.82%	74.39%	67.91%
All the features – (new word features)	94.44%	78.01%	69.68%
All the features	94.98%	81.66%	70.88%

6. Conclusions and future work

This paper investigates the problem of automatic prediction of focus word pair, and CRFs are introduced for the task. We demonstrate that CRFs outperform SVM experimentally. CRFs can take advantage of richer contextual information and predict a label for a single word by considering neighboring words and improve recall rate significantly. The performance of CRFs are also less affected by the imbalanced distribution between the positive and negative instances of the data set. To further improve the performance of the predictor, some new features are added to consider the specialty of different kinds, especially subset word pair, of focus word pair. Experimental results demonstrate the efficiency of the proposed method. In the future, we will incorporate acoustic information as features for the detection of focus word pair. Different discriminative sequence labeling models will also be investigated to obtain better performance. The results of focus word pair prediction will also be incorporated into our work [13] for emphatic and expressive speech synthesis.

7. Acknowledgements

This work is supported by the National Basic Research Program of China (2012CB316401 and 2013CB329304). This work is also partially supported by the Hong Kong SAR Government's Research Grants Council (N-CUHK414/09), the National Natural Science Foundation of China (61375027, 61370023 and 60805008) and the Major Project of the National Social Science Foundation of China (13&ZD189).

8. References

[1] T. Zhang, M. Hasegawa-Johnson, S.E. Levinson, "Extraction of pragmatic and semantic salience from spontaneous spoken English," Speech Communication, 2006, 48: 437-462.

[2] M. Rooth, "A theory of focus interpretation," Natural Language Semantics, 1992, 1: 75-116.

[3] S. Calhoun, J. Carletta, J. Brenier, N. Mayo, D. Jurafsky, M. Steedman, "The NXT-format Switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," Language Resources and Evaluation, 2010, 44: 387-419.

[4] S. Calhoun, "Information structure and the prosodic structure of English: a probabilistic relationship," Dissertation. University of Edinburgh, 2007.

[5] L. Badino, R. Clark, "Automatic labeling of contrastive word pairs from spontaneous spoken English," In: Proceedings of the 2008 IEEE/ACL Workshop on Spoken Language Technology, 2008: 101-104.

[6] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[7] J. Godfrey, E. Holliman, J. McDaniel, "Switchboard: Telephone speech corpus for research and development," In: Proceedings of the IEEE Conference on Acoustic, Speech, and Signal Processing, 1992, 1: 517-520.

[8] J. Francom, M. Hulden, "Parallel multi-theory annotations of syntactic structure," In: Proceedings of the Language Resources and Evaluation Conference, 2008: 2339-2443.

[9] M. Guegan, N. Hernandez, "Recognizing textual parallelisms with edit distance and similarity degree," In: Proceedings of the European Chapter for Association in Computational Linguistics, 2006: 281-288.

[10] T. Pedersen, S. Patwardhan, J. Michelizzi, "WordNet::similarity - Measuring the relatedness of concepts," In: Proceeding of the 19th National Conference on Artificial Intelligence, 2005: 38-41.

[11] "CRF++: Yet another CRF toolkit," http://chasen.org/~taku/ software/CRF++/, 2005-05-28/ 2014-03-20.

[12] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, C.J. Lin, "LIBLINEAR: A library for large linear classification," Journal of Machine Learning Research, 2008, 9: 1871-1874.

[13] F.B. Meng, Z.Y. Wu, J. Jia, H. Meng, L.H. Cai, "Synthesizing English emphatic speech for multimodal corrective feedback in computer-aided pronunciation training," Multimedia Tools and Applications, Springer, 2013: 1-27.