Selectional Preferences Based on Distributional Semantic Model

CHAO LI FUJI REN XIN KANG
Faculty of Engineering
Tokushima University
2-1 Minami Josanjima, 770-8506 Tokushima
JAPAN
c501447002@tokushima-u.ac.jp ren@is.tokushima-u.ac.jp kang-xin@is.tokushima-u.ac.jp

Abstract: In this paper, we propose a approach based on distributional semantic model to the selectional preference in the verb & dobj (direct object) relationship. The distributional representations of words are employed as the semantic feature by using the Word2Vec algorithm. The machine learning method is used to build the discrimination model. Experimental results show that the proposed approach is effective to discriminate the compatibility of the object words and the performance could be improved by increasing the number of training data. By comparing the previous method, the proposed method obtain the promising results with obvious improvement. Moreover, the results demonstrate that the semantics is an universal, effective and stable feature in this task, which is consistent with our awareness of using words.

Key–Words: Selectional preferences, Distributional semantic model (DSM), Lexical semantics, Word2vec algorithm

1 Introduction

For humans, the ability of association and inference plays an important role in language learning and organization. Resnik presented the earliest work in this field [1][2]. Following Resnik's work, "Selectional Preference" is used to describe this task. For instance, when people learn the usage of verb 吃 chi "eat", they primarily learn some examples like 吃苹果 chipingguo "eat apples" and 吃桔子 chijuzi "eat oranges", and then infer that 吃香蕉 chixiangjiao "eat bananas" is a correct expression, because we know that apples, oranges and bananas could be eaten. For another instance, given a new word 冰桶挑战 bingtongtiaozhan "Ice Bucket Challenge", we know how to associate this word in our language, like 参与冰桶挑战 canyubingtongtiaozhan "take on the ice bucket challenge", based on its semantic meaning. Therefore, to make computers learn and use language effectively and flexibly, we need to develop the ability of association and inference on words in the computer, based on the semantics of words.

In a sentence generation, for example, the feasibility of semantics in words is very important for generating readable sentences. Gözde Özbal et al. (2013) presented the typical errors in their generated sentences "Unscrupulous doctors smoke *armored units*", by specifying the keywords and other conditions in a creative sentence generation system [3]. The direct-object *armored units* is considered infeasible to the rest of the sentence. In another example, "A pleasant tasting, a *heady* wine", "*heady*" is not a property of "wine". The reason of these errors is that their system do not refer the plausibility of semantic between words. Moreover, selectional preferences was widely used in many natural language processing tasks, such as word sense disambiguation [4], dependency parsing [5], semantic role labeling [6].

In this study, we learn a model for selectional preference task in Chinese based on the semantic information, and try to relieve its dependency on a restricted vocabulary in a corpus. Specifically, we want to recognize "eat apples" as a correct expression and classify "eat problem" as an incorrect expression. To simplify the experiments, we focus on the verb & dobj relation in this paper, and leave the compatible discrimination for other kinds of word relations in the future work. The reason of studying the verb & dobj relation is that verbs are important for interpreting meanings in sentences. For example, "eat an apple" or "buy an apple " takes more specific meaning than a single word "apple ". Moreover, Capturing this aspect in the verb & dobj relation can improve the plausibility of sentence generated by machine [3].

In the experiments, we employ a supervised classifier for selectional preference task between a word and a specified verb in the verb & dobj relation. For example, by training the classifier with compatible samples like 吃苹果 chipingguo "eat apples" and 吃桔子 chijuzi "eat oranges", we expect it to infer 吃香

蕉 chixiangjiao "eat bananas" as a correct expression. For the semantic feature, we employ the Word2Vec algorithm to generate vector representations of words. Experimental results suggest that our approach is effective to discriminate the compatibility for words in the verb & dobj relation, and the performance can be further improved by increasing the number of training examples. Further, in order to investigate the impact of distributional semantic model, the DSP (Discriminative Selectional Preference) method proposed by Bergsma et al. [7], which also represented each sample as a numeric vector, is employed to make comparison with the proposed method.

The reminder of this paper is organized as follows. Section 2 reviews some related work on the selectional preferences. Section 3 introduces the details of proposed approach. The experimental results and discussion are presented in section 4. Finally, section 5 concludes this study with future work.

2 Related work

2.1 Selectional Preference

Selectional preferences has long been considered a fundamental problem in computational semantics which could discriminate which arguments are plausible for a particular predicate [7]. Many methods have been presented to study selectional preferences problem. Ritter et al. concluded Four categories among these methods: class-based methods, similarity-based methods, discriminative methods, and generative probabilistic models [8].

In 1996, Resnik presented the first class-based method, which identified the plausibility of predicate and argument based on WordNet [2]. In WordNet, words are classified in hierarchical layers. In alternative, Pantel presented a method in which the classes of words were generated automatically based on clustering [9]. Following Rensik, many other class-based method were presented. Li et al. and Jia et al. investigated selectional preferences in Chinese language based on HowNet[1] [10][11]. HowNet is similar to WordNet in Chinese.

In order to overcome the limitation of the coverage of class-based method, Erk employed the similarity-based model for selectional preferences [12]. In 2013, a random walk model was employed for selectional preferences [13].

Ritter et al. employed a latent Dirichlet allocation method to conduct selectional preferences task, and referred their method as a generative probabilistic model [8]. In 2012, Jang and Mostow introduced the PONG (Part-Of-Speech N-grams) method to conduct selectional preferences for different relations with probability based on part-of-speech and N-grams model.

In 2008, Bergsma et al. presented the DSP method for selectional preferences based on discriminative method [7]. In their method, each word was a sample. Each sample had several features, and each feature was assigned with a numeric value. And then, each word/sample was represented by a vector of numeric values. Three kinds of features were employed: verb co-occurrence, string-based features and semantic classes. The verb co-occurrence features were probabilistic values. The string-based features were frequent and boolean values. The semantic classes based on CBC (Clustering based on Committee) method [9][15]. Finally, the machine learning algorithm, SVM (support vector machine), is used to build the discriminative model.

In this paper, the proposed method will compare with Bergsma et al.'s method. The difference is that the features are extracted by the novel method in proposed method, which base on the distributional semantic model. In other words, the word vector is used to represented each word/sample.

2.2 Word Vector

Word Vector is a distributional representations of words by using the continuous bag-of-words and skip-gram architectures. The input of this method is a corpus. The output is the word vectors in which each vector represent the semantic of a word. Word2Vec[2], published by Google in 2013, is an implementation of distributed representations of words [16, 17, 18]. The dimensional size of the vector could be set at beginning of training the word vector model.

2.3 Dependency Grammar

Dependency is the notion that linguistic units, e.g. words, are connected to each other by directed links. The (finite) verb is taken to be the structural center of clause structure. All other syntactic units (words) are either directly or indirectly connected to the verb in terms of the directed links, which are called dependencies[3]. In this paper, the Standord parser is used to retrieve the "dobj" dependency relation between verb and other words from corpus [19], in order to reduce the workload of preparing the experimental data.

[1] http://www.keenage.com

[2] https://code.google.com/p/word2vec/

[3] https://en.wikipedia.org/wiki/Dependency_grammar

3 Approach

We propose a novel approach for selectional preference task in Chinese, based on the structural and lexical semantics information of words, and try to relieve the dependency on restricted vocabularies. Our task could be viewed as a word similarity problem and a binary classification task with specific verbs. The three main factors in the proposed approach are data, word representation, and learning.

3.1 Preparing Data

Fudan corpus[4][5] as a Chinese corpus from Fudan University is employed in our work, for training and evaluating our model of selectional preference. We extract examples of the verb & direct object relation with Stanford parser[6] from this corpus for reducing the workload, and manually revise the incorrect examples for improving the annotation. For example, dobj(提高, 效率) dobj(tigao, xiaolü) "dobj(improve, efficiency)" is a correct example, while dobj(提高, 总产量) dobj(tigao, zongchanliang) "dobj(improve, total output)" is an incorrect example because "output" rather than "total output" is the direct object of "improve". We revise this examples by dobj(提高, 产量) dobj(tigao, chanliang) "dobj(improve, output)" in our approach. Besides, we create negative examples by randomly selecting the direct objective words, like dobj(提高, 提高) dobj(tigao, tigao) "dobj(improve, improve)". We also correct the segmentation errors in the parsing result.

3.2 Word Representation

Generally, there are many words which are compatible with a specific verb. Our assumption is based on the similarity between a new word and the observed words with respect to a specified verb for selectional preference task. We employ the structural feature and the semantic feature for evaluating the word similarity.

For the structural feature, we employ the separate Chinese characters and the number of Chinese characters in the words. These features are designed by the fact that Chinese words with similar morphology have the similar usage. For example, the direct object in (提高, 速度) (tigao, sudu) "(improve, speed)" and (提高, 精度) (tigao, jingdu) "(improve, accuracy)" consist of the same character "度".

For the semantic feature, we employ the distributed representation of words generated by the Word2Vec algorithm. Word2Vec is a unsupervised feature learning algorithm based on a recurrent neural network, which learns vectors of float numbers for representing the semantic meanings of different words. We train a Word2Vec model on the SogouT corpus[7] [21], to get vectors for 421609 words, by using the gensim which is python package with the implementation of Word2Vec algorithm.

We use the structural feature and semantic feature separately, and also combine these features, for evaluating the word similarity and for developing selectional preference model.

3.3 Compatibility Discrimination Learning

To learn a model for selectional preference, we employ the machine learning algorithms provided by scikit-learn package[8]. We train one classifier for each verb, and select large amount of examples of the compatible and incompatible direct objects as the training and testing data. For instance, in dobj(提高, 效率) dobj(tigao, xiaolü) "dobj(improve, efficiency)", the word 效率 xiaolü "efficiency" is used as an example. For all examples, we extract their structural and semantic features as described above.

4 Experiment

We select five verbs and construct the training and testing data with the compatible and incompatible direct objects. And then, in order to investigate the impact of distributional semantic model for selectional preferences, two groups of experiments are conducted. In the first group of experiments, the logistic regression classifiers are trained for each verb, based on the structural and semantic features as well as the combined features respectively, for selectional preference. In the second group of experiments, the proposed and Bergsma et al.'s approach DSP are compared by training the SVM (Support Vector Machine) classifiers. Macro-averaged Precision, Recall, and F_1-score are employed to evaluate classification results, and analyze the performance improvements with respect to different features and different training data sizes.

4.1 Dataset and tools

Two datasets, several NLP (natural language processing) and machine learning tools are employed in the experiments.

[4]http://www.datatang.com/data/44139

[5]http://www.datatang.com/data/43543

[6]http://nlp.stanford.edu/software/lex-parser.shtml

[7]http://www.sogou.com/labs/resource/t.php

[8]http://scikit-learn.org

4.1.1 Dataset

We employ Fudan text classification corpus for developing the training and testing examples for five verbs. This corpus consists 19,637 documents about 187MB. The selected verbs include 提高 tigao "improve", 进行 jinxing "progress", 作为 zuowei "act as, used as, etc.", 得到 dedao "get, obtain, receive, etc.", 建立 jianli "establish etc.". We extract the verb & direct object pairs in the direct object relation from dependency parsing results which generated by the Stanford parser. These pairs are manually examined by three human experts. Finally, we select 1000 positive examples and 1000 negative examples, for each verb.

We employ a subset of SogouT corpus, which has been released by Sogou.com, for developing the Word2Vec model and extracting the features for DSP method. The SogouT corpus consists 1.9GB Chinese news collected from the Internet.

4.1.2 Tools

Stanford parser is a program that works out the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb[9]. In this paper, the Chinese dependency parser is used to extract samples in dobj and nsubj relation from the corpora, which is used as the training, testing data and used to extract features for DSP approach [22].

Gensim[10], which is an implementation of Word2Vec algorithm in python language, is employed to obtain a vector representation of words that implicitly represent some lexical semantic information [20] in the experiments.

Scikit-learn (formerly scikits.learn) is an open source machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, logistic regression, naive Bayes, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy[11]. The function of logistic regression and Linear SVM classifier are employed in this paper. To evaluate the results, we also use the functions to compute the precision, recall and cross validation provided by Scikit-learn[12].

S-Space Package, which is in Java language, is a collection of algorithms for building Semantic Spaces as well as a highly-scalable library for designing new distributional semantics algorithms[13]. In this paper, S-Space is used to reproduce CBC algorithm in DSP approach.

4.2 Results and Analysis

To examine selectional preference models based on distributional semantic model, we have randomly extract five groups of training examples, with 200, 400, 800, and 1000 examples in each group. We repeat each random extraction for 10 times for the model training. And the evaluation scores are averaged over 10 models.

4.2.1 Results based on distributional semantic model

Fig.1 shows the macro-averaged Precision, Recall, and F_1-score, for five verbs exploited in the experiments. We abbreviate the semantic feature generated by the Word2Vec algorithm as "w2c", the structural feature as "form", and the combination of these two features as "comb". The sizes of training data are represented on the x-axis.

Our results suggest that the selectional preference based on semantic feature, could always be improved by feeding with more training examples. The macro-averaged Precision, Recall, and F_1-score get consistent increments by training on larger data sets. This demonstrates that semantic feature is effective for discriminating the compatibility between words in the verb & dobj relation, which is also consistent with our awareness.

Training with the structural features does not obtain obvious improvements with larger training data sets. By analyzing the classification results for each verb, we find that the structural feature has effectively increased the Precision scores for 提高 tigao "improve", and 建立 jianli "establish". The reason is that the direct object words for these two verbs show very similar morphology in the used data set. This also causes the classifier based on combined feature rendering a higher macro-averaged Precision than the semantic feature based classifier. On average, we find the semantic feature is better than the structural feature and even the combined feature, which renders a universal feature and a stable performance.

4.2.2 Results based on comparison with DSP method

Fig.2 shows the comparison of the proposed approach and DSP in macro-averaged Precision, Recall, and F_1-

[9] http://nlp.stanford.edu/software/lex-parser.shtml
[10] https://radimrehurek.com/gensim/
[11] http://en.wikipedia.org/wiki/Scikit-learn
[12] http://www.scikit-learn.org
[13] https://github.com/fozziethebeat/S-Space

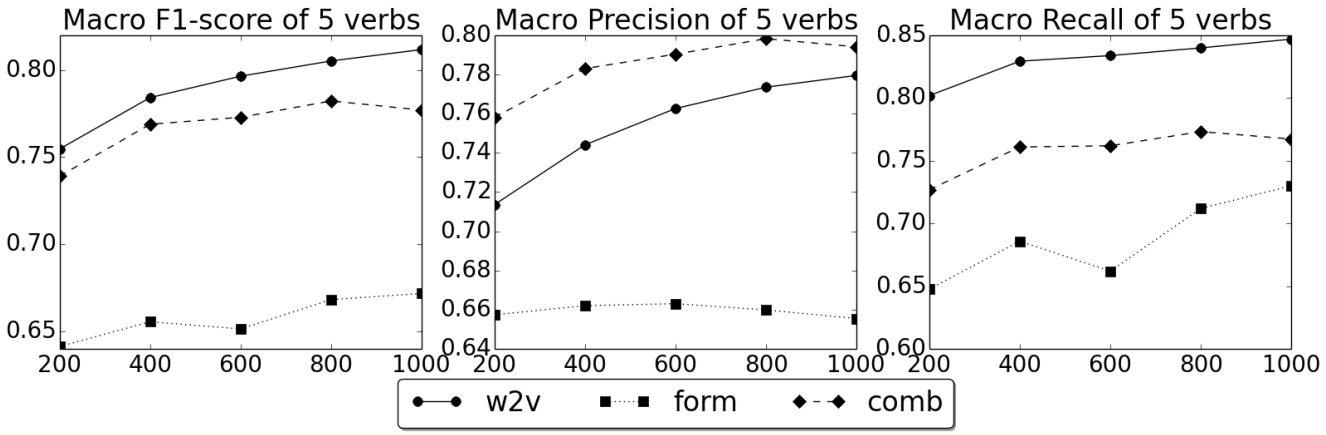

Figure 1: The Macro F_1-score, Precision and Recall of 5 verbs with logistic regression classifier. In the legend, "w2v" is the semantic feature represented by the Word2Vec algorithm. "form" is the structural feature. "comb" is the combination of these two kinds of feature. The horizontal axis indicates the amount of training data.

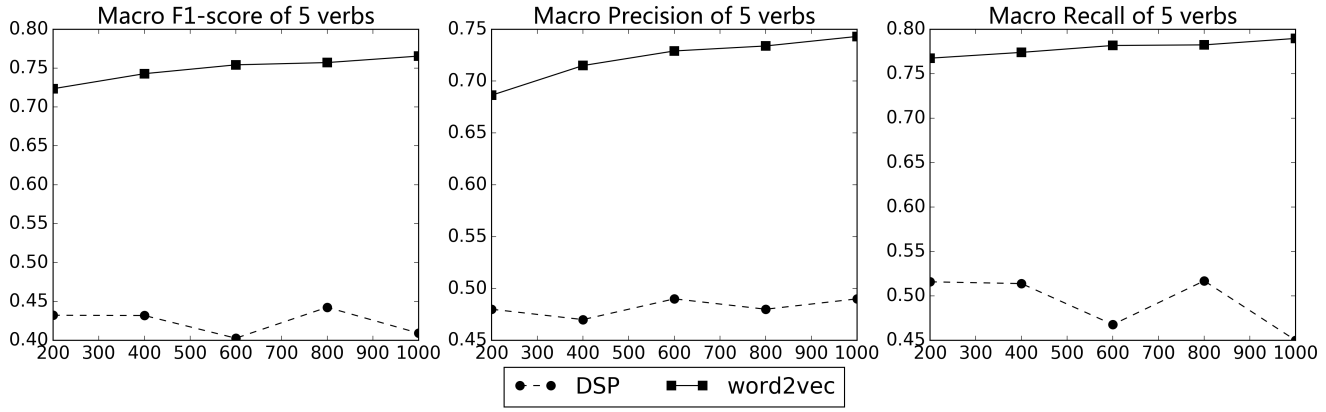

Figure 2: The comparison of the proposed approach and DSP method with SVM classifier. In the legend, "Word2vec" is the semantic feature represented by the Word2Vec algorithm. "DSP" is the results of DSP method. The horizontal axis indicates the amount of training data.

score with, for five verbs exploited in the experiments.

The DSP method used the SVM classifier. Therefore, we also use SVM classifier in the comparison. Specially, the results are a little different with figure 1, because selecting samples randomly make the training and testing data different in these two groups of experiments. Although the data has changed, the results show the consistent tendency when the number of training data is growing.

The results suggest that using distributional semantic model could outperform the DSP method for selectional preference. And the results could not be improve with more training data by using the DSP method. The DSP method performed well in Bergsma et al.'s paper, resulting in 0.60 marco-precision and 0.71 recall, with a large set of data which is larger than us. Therefore, the results of the DSP method are not good in our experiments. However, the proposed approach could obtain the promising results with smaller size of data that used in the DSP method.

In details, the clustering feature in DSP method, which is based on CBC algorithm, spent a lot of time in our experiments. At the beginning, there are more and 70 thousands words in the corpus. By using the S-space tools to implement the CBC algorithm, the program still kept running after 2 weeks. Considering the efficiency, we reduce the number of words to about 40 thousands by raising the threshold which is the frequency of occurrence. This threshold was set as 20, and we only extract the words in "nsubj" and "dobj" relation. With the new threshold, it spent about 1 week to obtain the results of CBC algorithm. The String-based features are not employed in the experiments, because our experiments only focus on Chi-

nese words which do not have those String-based features presented in the DSP method.

5 Conclusion

In this paper, we proposed a semantic and structure based approach to discriminate the compatibility of words. To simplify the experiments, in this paper we only focused on the compatibility of verbs and their direct objects. We employed the Word2Vec algorithm to extract the distributed representations of words as their semantic feature, and integrated the forms of words as the structural feature. In the experiments, we examined the separate features as well as the combination of two features in compatibility discrimination for five specified verbs. Subsequently, by comparing of the proposed method and the DSP method, the results suggested that using distributional semantic model could improved the performance for selectional preference. Our results suggested that the proposed approach is effective to discriminate the compatibility of words, and the performance could be further improved by feeding more training examples for the compatibility classifiers. The results also demonstrated that the semantic feature is an universal, effective, and stable feature for selectional preference task.

In the future work, we will integrate more words to evaluate the proposed method subsequently, and extend the proposed approach to other verbs by computing the similarity between these verbs. We also are interested in evaluating the proposed method with other dependency relation such as "subj" (subject of a verb) and "amod" (attributive adjectival modifier).

Acknowledgment

This research has been partially supported by JSPS KAKENHI Grant Number 15H01712.

References:

[1] Philip Resnik, Semantic classes and syntactic ambiguity, In *Proceedings of the workshop on Human Language Technology*, 1993, pp. 278–283

[2] Philip Resnik, Selectional constraints: An information-theoretic model and its computational realization, *Cognition*, 1996, pp. 127–159

[3] Gözde Özbal, Daniele Pighin and Carlo Strapparava, BRAINSUP: Brainstorming Support for Creative Sentence Generation, In *ACL* 2013, 2013, pp. 1446–1455

[4] Philip Resnik, Selectional preference and sense disambiguation, In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, 1997, pp. 52–57

[5] Guangyou Zhou, Jun Zhao, Kang Liu and Li Cai, Exploiting web-derived selectional preference to improve statistical dependency parsing, In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 1556–1565

[6] Daniel Gildea and Daniel Jurafsky, Automatic labeling of semantic roles, *Computational linguistics*, 2002, vol. 28(3), pp. 245–288

[7] Shane Bergsma, Dekang Lin and Randy Goebel, Discriminative learning of selectional preference from unlabeled text, In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 59–68

[8] Alan Ritter, Mausam and Oren Etzioni, A latent Dirichlet allocation method for selectional preferences, In *ACL* 2010, 2010

[9] Patrick André Pantel, Clustering by committee, Ph.D. thesis , University of Alberta, Edmonton Alta., Canada, 2003

[10] Bin Li, Xiaohe Chen and Xuri Tang, An Investigation of Chinese Selectional Preference Based on HowNet, Genetic and Evolutionary Computing (ICGEC), 2010 Fourth International Conference on, 2010, pp. 236–239

[11] Yuxiang Jia, Hongying Zan and Ming Fan, Inducing Chinese Selectional Preference Based on HowNet, Computational Intelligence and Security (CIS), 2011 Seventh International Conference on, 2011, pp. 1146-1149

[12] Katrin Erk, A simple, Similarity-based model for selectional preferences, In *ACL* 2007, 2007

[13] Zhenhua Tian, Hengheng Xiang, Ziqi Liu and Qinghua Zheng, A Random Walk Approach to Selectional Preferences Based on Preference Ranking and Propagation, In *ACL* 2013, 2013

[14] Hyeju Jang and Jack Mostow, Inferring Selectional Preferences from Part-Of-Speech N-grams, In *EACL* 2012, 2012

[15] Patrick Pantel and Dekang Lin, Discovering word senses from text, In *KDD* 2002, pp. 613-619

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space, In *Proceedings of Workshop at ICLR* 2013, 2013

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, Distributed Representations of Words and Phrases and their Compositionality, In *Proceedings of NIPS* 2013, 2013.

[18] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, Linguistic Regularities in Continuous Space Word Representations, In *Proceedings of NAACL HLT* 2013, 2013.

[19] Danqi Chen and Christopher D Manning, A Fast and Accurate Dependency Parser using Neural Networks, In *Proceedings of EMNLP* 2014, 2014.

[20] Radim Řehůřek and Petr Sojka, Software Framework for Topic Modelling with Large Corpora, In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

[21] Yiqun Liu, Min Zhang, Rongwei Cen, Liyun Ru and Shaoping Ma, Data cleansing for web information retrieval using query independent features,Journal of the American Society for Information Science and Technology, 2007, **58** (12) pp. 1884–1898.

[22] Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning, Discriminative Reordering with Chinese Grammatical Relations Features. In Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation, 2009.