Dialog-Act Recognition Using Discourse and Sentence Structure Information

Keyan Zhou[1,2] Chengqing Zong[1,2]
[1]*NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190*
[2]*China-Singapore Institute of Digital Media, 25 Heng Mui Keng Terrace, Singapore*
{kyzhou, cqzong}@nlpr.ia.ac.cn

Abstract

Automatic recognition of Dialog-act (DA) is one of the most important processes in understanding spontaneous dialog. Most existing studies have been working on how to use various classifying methods in DA recognition; meanwhile, less attention has been paid to feature selection specifically. This paper introduces several textual features for DA recognizing, and proposes a novel usage for sentence structure features. Especially, this paper investigates the effect of discourse structure features in DA recognition, which are little studied before. The experimental results on both Chinese corpus and English Corpus show the selected features and feature combination rules significantly improve the overall performance. The accuracy of DA recognition rises from 77.05% to 88.21% on Chinese corpus, and from 59.08% to 64.92% as well on English corpus.

1. Introduction

Dialog-act (DA), defined as the meaning of an utterance at the level of illocutionary force (Austin, 1962), reflects the intention of a speaker and the effect of a dialog utterance. DA has been widely used in language and speech processing, such as speech recognition (Dhillon *et al.*, 2004), spoken dialog system (Walker and Passonneau, 2001), summarization (Stolcke *et al.*, 2000), and spoken language translation (Reithinger and Maier, 1995; Sridhar *et al.*, 2008).

Over the few past decades, several projects on studying DA have been pursued. Verbmobil project (Reithinger and Maier, 1995), Switchboard telephone speech corpus project (Godfrey *et al.*, 1992), and ICSI meeting recorder project (Morgan *et al.*, 2001) are three noted projects. The Verbmobil project first used DA into speech-to-speech translation task. The latter two collected and annotated influential public corpora SWBD (Jurafsky *et al.*, 1997) and ICSI-MRDA (Janin *et al.*, 2003) respectively.

A variety of classification methods have been applied in DA recognition, including some traditional models, such as maximum entropy model and decision tree based classifier, as well as up-to-date machine learning techniques, such as Support Vector Machine (SVM) and graphical model. However, no much research has been done on feature selection and combination in the area of DA classification.

In this paper, we introduce several textual features at both sentence level and discourse level. The effect on feature selection and combination has been specially studied. Experimental results are given on both Chinese corpus and English corpus. The large improvement of overall accuracy proves the effectiveness of the selected features and feature combination rules.

The remainder of this paper is organized as follows. Section 2 introduces the related work and analyzes the features used in existing approaches. Section 3 gives our motivations and details on feature selection. Experimental results are presented and analyzed in Section 4. Finally, Section 5 draws some conclusions and outlines the future work.

2. Related Work

Automatic recognition of DA is a typical classification task. Basically, the features used for DA classification may be obtained from separated sentence or whole discourse, known as sentence structure and discourse structure.

The following models are well-known in using sentence structure features: maximum entropy method (Ang *et al.*, 2005), decision tree based classifier

(Stolcke *et al.*, 2000), and SVMs (Surendran and Levow, 2006). A few studies have been done on feature analysis at sentence level. Kral *et al.* (2006) proposed a sentence structure definition. Verbree *et al.* (2006) got a major improvement on different corpora using smart features. Features used in the above mentioned work are listed in Table 1 with author abbreviation.

Table 1. **Sentence level features used in previous work with author abbreviation**

Features	A.	M.	S.	K.	V.
Unigrams		√	√	√	√
Bigrams	√		√		√
Trigrams			√		√
POS					√
Frequency			√		
Position	√			√	
Length	√				√
Sequence		√			

Having investigated on the characteristics of the sentence structure features, we further classify the features into the following three types:

(1) **Word level:** consists of word units including unigrams, bigrams, trigrams, etc.

(2) **Syntax level:** consists of labels such as part-of-speech (POS), chunks recognized as base noun phrase/verb phrase (BNP/BVP) and so on.

(3) **Restraint information:** refers to the syntactic or semantic constraints on units, including their position in utterance, weight of each feature in the two types mentioned above, utterance length, the frequency of units, and restraints on specific ambiguous words (known as word sense disambiguation (WSD)).

As shown in Table 1, the word level features are commonly used in previous studies. Extraordinarily, bigrams used in (Ang et al., 2005) only involve the one at the beginning or at the end of an utterance; similarly, in (Surendran and Levow, 2006), trigrams refer to those appearing at least twice in the training set.

Syntax level features are less discussed comparatively. Only Verbree *et al.* (2006) made use of POSs. Other syntax level features like Chunks have never been studied in the previous work.

As Jurafsky *et al.* (1998) mentioned, the common view considers words and phrases as the strongest features in DA classification. Comparably, restraint information has always been considered as a secondary factor.

Discourse structure can be modeled by DA based *n*-gram model, hidden Markov model (HMM) (Stolcke *et al.*, 2000; Surendran and Levow, 2006), Graphical model (Ji and Bilmes, 2005) and Markov Decision Process (MDP) model (Zhou *et al*., 2008). Features used in these work are mostly previous speaker, DA sequence. Other discourse structure such as adjacency pairs (APs) and topics are rarely discussed.

The APs are paired utterances, defined as one kind of sociolinguistic facts about conversation structure, which is a reflection of dialog structure (Levinson, 1983). The APs describe how participants might expect one type of dialog units to be responded to by another (Jurafsky *et al.*, 1997), such as question-answer, greeting-greeting, and so on. Galley *et al.* (2004) proved that AP can help in identifying whether an utterance expresses an agreement or disagreement. They also believed that AP would be useful in other computational pragmatics research such as DA classification. However, no experimental work has been reported yet.

Bangalore *et al*. (2006) gave a DA discourse structure model in range of same topic for doing topic segmentation. In their experiments, DA feature is not helpful for topic labeling. Unfortunately, they did not give any further discussion. Actually, the relationship between topic and DA differs in different topics. For example, in topic "Order-Item" in hotel-reservation domain, which includes booking rooms and other activities, the DA "Imperative" occurs more frequently than others. Meanwhile, in some other topics, such as "Furnishing", "Time", the constraints between topics and DAs are less tight.

In summary, feature selection and combination for DA classification are still open questions. Especially, only a few works have made use of both sentence and discourse structures so far.

3. Our Motivations and Methods

3.1. Sentence Structure Features

We have proposed a novel division on sentence structure features Section 2, which includes word level, syntax level, and restraint information. Features used in this paper are listed in Table 2.

(1) Word Level. For word level, we specifically refer to the unigram, bigram, and trigram. Considering the problem of data sparseness after $n \geq 3$ (Surendran and Levow, 2006), we will also research on relations between performance and frequency of *n*-gram.

(2) Syntax Level. We get POS information using ICTCLAS Tagger (http://ictclas.org/) for Chinese

Table 2. **Our sentence structure feature set**

(1) Word Level	Unigrams (UNI) Bigrams (BI) / Trigrams (TRI)
(2) Syntax Level	POS (POS) Base NP(BNP)
(3) Restraint Information	Position (PST) / Weight (WT) Utterance Length (UL) Word Sequence(WS) Frequency (FQ) / WSD (WSD)

corpus, and Stanford POS Tagger (http://nlp.stanford.edu/software/tagger.shtml) for English corpus. POS is appended to unigram as a supply to individual words.

BNP information is automatically annotated based on (Xu *et al.*, 2006). Once a chunk is labeled as BNP, it will be replaced by a tag regardless of its content.

(3) Restraint Information. Position information concerns the relationship between the ability a word acts to DA and its position in the utterance. At the DA level, words closer to the beginning and the end of a sentence are more important than the others.

For unigram, in an utterance $U = u_1 \ldots u_i \ldots u_n$, $D(u_i)=|i-k|$, $k=(n+1)/2$ denotes the relative position of word u_i. $D(u_i)$ increases symmetrically as u_i appears nearer to the end or the beginning of the utterance. We introduce an exponential weight α on $D(u_i)$ that controls the contribution of position feature to the unigram feature. Therefore, position information of unigram can be symbolized as (a). For position restraint of bigram, we simply consider the first and last bigrams of an utterance as Ang *et al.* (2005) did.

$$P(u_i) = |i-k|^{\alpha}; \quad k = (n+1)/2 \quad \text{if } i = k, P(u_i) = (|i-k-1|^{\alpha})/2 \qquad \text{(a)}$$

Using multiple knowledge sources, we need a restraint to make sure the combination is effective. Therefore, we assign weights to different units at word level based on their length. We believe longer units contain more information. The specific value will be given empirically in the experiments.

Frequency is the direct representation of data sparseness, which also reveals the importance of a feature in a sense. For reducing the noise brought by data sparseness, data filtering based on frequency will make the training more efficient.

The average utterance length of Chinese dialog corpus is about 7 Chinese words including punctuation, since sentences are cut to make sure that an utterance only contains one specific DA label. Utterances with one single word are always "backchannel" or "accept" of speakers which belong to DA "s". Therefore, we assign a special sort of *UL*=1. For utterances with more than one words, we roughly divide them into $2 \le UL \le 10$ and $UL > 10$.

Ambiguous words are defined as words with multiple senses. Ambiguous words commonly exist in spoken dialogs, especially in the Chinese corpus. Particularly, in dialogs, the sense of an ambiguous word is always related to a specific DA. For example, "还是(or / is still)" can be a query or a statement, as shown in the following examples:

1) 你 是 要 单人 间，还是 双人 间 ? (Do you want a single room, *or* double room ?)

2) 这样 还是 比较 好. (It *is still* good.)

The sentences with these ambiguous words are always misclassified. Thus, we introduce the concept of WSD that contains several restraints for ambiguous words. In this paper, these restrains include word position, word sequence (WS) nearby ambiguous words, and punctuation of the utterance. When an ambiguous word is detected, it will be labeled with "B", "M" or "E" based on its position in the utterance. Then, punctuation of the utterance will be appended to the ambiguous word. The word sequence nearby ambiguous word will be considered only if *length*[WS]≤5 and FQ≥2 in corpus.

3.2. Discourse Structure Features

According to (Grosz and Sidner, 1986), discourse structure supplies the information for conversational participants, so that they can determine how an individual utterance fits in with the discourse.

The topic and APs are two basic representations of discourse structure. A dialog is composed of several topics. Each topic contains a DA sequence, which is an abstract of specific utterances. AP describes the restriction between specific kinds of DAs within the range of a topic. This paper will give experiments and focus on revealing how topics and APs affect in DA classification.

4. Experiment and Analysis

4.1. Data and Labels

We use Chinese human-human dialogs (CH corpus) (Zhou *et al.*, 2008) in the domain of hotel-reservation. There are 174 dialogs, consisting of 6,208 utterances, which are transcribed from conversational telephone speech and manually corrected. The average utterance length is 7 words, about 12 characters. The dialogs are labeled with DAs, APs, and topic manually (Zhou *et al.*, 2008).

The APs contain the following relationships: question-answer, greeting-greeting, offer-acceptance, and apology-downplay. Topics contain 9 categories: Greeting <G>, Price <P>, Furnishing <F>, Time <T>, Contact-Information <C-I>, Check-Out <C-O>, Order-Item <O-I>, Ending <E>, and Others <O>.

4.2. Classifiers

As features belong to sentence structure or discourse structure are adaptive to distinct models, we employ an SVM classifier and an MDP model to process the two sorts respectively.

SVM is employed to classify DAs based on sentence structure of individual utterance. We use a well-known SVM tool libsvm-2.84 (http://www.csie.ntu.edu.tw/~cjlin/libsvm) for its convenience and utility. MDP defined as a tuple $\{S, A, T, R\}$ is utilized to predict the DA sequence based on discourse structure. S stands for dialog state, which is composed of speaker and DA history in baseline system (DSB). The more discourse structure information including TP and AP is then added into dialog state for comparative experiments.

Finally, the predicting results of MDP will be trained as features of SVM classifier as Zhou *et al.* (2008) did.

4.3. Baseline Features

The baseline features are unigrams (UNI) including unigram words, and punctuations for SVM, and DSB consists of speaker and the DA history for MDP.

We evaluate the performance of baseline features in CH corpus. The results are presented in Table 3 measured by accuracy with 5-fold cross validation.

$$Accuracy = \frac{\#\text{number of correctly predicted DAs}}{\#\text{total number of predicted DAs}} \times 100\%$$

We get an accuracy of 77.05% by using UNI in SVM, and an improvement of 1.06% by integrating MDP prediction using DSB.

4.4. Effect of Sentence structure Features

The sentence structure features are added into SVM one by one. The results are shown in Table 3. The abbreviations of features can be seen in Table 2. In feature combinations, '+' represents adding features, while '_' represents adding restraints.

Using BI and TRI alone get lower accuracies than baseline UNI. Note that adding BI to UNI hurts the performance seriously from 77.05% to 75.85%. In contrast, after adding PST constraints to UNI (as α=1/2 in formula (a)) and BI, the accuracy rises to 77.90%.

Considering data sparseness, we introduce constraint FQ. Given n[5]=[1, 5, 20, 100, 200], for each i=1,…,5, UNIs, BIs, and TRIs are selected only if their FQ≥n[i]. The accuracy and FQ curves are shown in Figure 1. The curve of UNI shows a steady improvement when FQ increases. The accuracy of BI gets the maximum as FQ = 100, while TRI gets the maximum as FQ = 20.

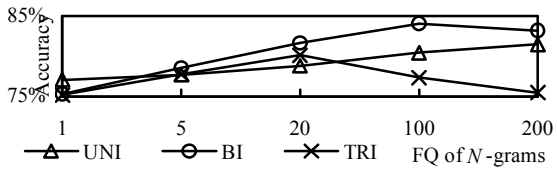

Figure 1. **Influence of data frequency**

We also experiment on combination for UNI, BI and TRI, only a few typical results are listed in Table 3. (UNI+BI)_FQ200 gets a better accuracy of 87.11% than other combinations. Adding TRI_FQ100 hurts the accuracy, which challenges the common view that longer units contain more information. However, longer units also result in more sparseness. In the following test, we adjust WT for UNI and BI, the accuracy first goes up and then down when WT charges from 1:1 to 0.5:1. The best result 87.21% is got when WT = 0.85:1, which confirms that BI is more significant than UNI. We also conclude that UL does help in DA classification.

In syntax level, a slight improvement occurs when POS is appended to UNI. Meanwhile, BNP improves the accurate more effectively. In CH corpus, 10,457 BNPs out of 44,059 UNIs are recognized, which greatly reduces the size of feature space.

The experimental results reveal the following two deficiencies of features: sparseness and redundant feature space. For the first problem, we use FQ restraint to remove sparse features. For the latter one, we extract BNP chunks to obtain a refined feature set.

Finally, we investigate several ambiguous words. As shown in Table 4, the disambiguation doesn't work well as we expected. On the whole, our restraint rules seem to be a little weak comparing with complicated situations of ambiguous words. To improve the performance of WSD, some semantic level restrains might be needed.

4.5. Discourse structure Features

Table 3. **Accuracies(%) of DA recognition in CH corpus**

Feature Structure	Feature Combination	Accuracy (%)
Baseline Features	UNI / UNI+DSB	77.05 / 78.11
Sentence Structure	BI / TRI	75.30 / 75.20
	UNI+BI / (UNI+BI)_PST	*75.85* / 77.90
	(UNI+BI)_FQ200	87.11
	(UNI+BI)_FQ200+TRI_FQ100	*86.45*
	(UNI+BI)_FQ200_WT	87.21
	(UNI+BI)_FQ200_WT_UL	**87.62**
	UNI+POS / UNI+BNP	77.33 / 80.13
	(UNI+POS+BNP)_FQ200	83.26
	(UNI +POS+BNP+BI)_FQ200_WT _UL	87.85
	(UNI +POS+BNP+BI)_FQ200_WT _UL_WSD	**87.96**
Discourse Structure	UNI+DSB+TP3 / UNI+DSB+AP	78.36 / 80.54
Best Performance	(UNI+POS+BNP+BI)_FQ200_WT_UL_WSD+TP3+AP	**88.21**

The discourse structure features are added into DSB of the MDP model. Results are shown in Table 3.

We get a large improvement of accuracy from the baseline 77.05% to 88.21%. Noticeably, we use TP3 instead of entire 9 TPs in Table 3. That's because we find that only a few topics give positive results. The contribution of each TP is given in Table 5. Through investigation of corpus, we find that the improvement TP3 brought is closely related to a few specific DA tags. For example, 72.2% of "qh" and 61.4% of "qo" appear in TP <P> and <O-I>, while "is" in TP <C-I> and TP <O-I> takes 67.1%. It is obvious that the TP can greatly improve the recognition accuracy of these DAs. Comparatively, the other kinds of DAs do not benefit from TP.

Results show AP is an effective feature. This is because AP is the discourse structure description closest to DA. AP directly reflects the relationship between a pair of DAs, which can be considered as an essential structured DA sequence.

4.6. Application to a Public Corpus

We also apply our features to the SWBD corpus (Jurafsky *et al.*, 1997). The SWBD is a public corpus of conversational telephone speech with DA annotation. The tag set in this paper contains 42 out of the original 220 DA-labels, similar to Stolcke *et al.* (2000) and Verbree *et al.* (2006). We choose 220 dialogs (40,382 utterances) randomly and perform 5-fold cross validation as we did in CH corpus.

As shown in Table 6, features act similarly as in CH corpus. One exception is that UNI +BI gets better accuracy (59.21%) than UNI itself (59.08%), which differs from their performances in CH corpus. To find the reason, we did a comparable test on a smaller size of SWBD corpus. When we reduce utterances to 10,000, the accuracies of using UNI+BI and UNI itself are 53.41% and 54.95% respectively. It is clear that the efficiency of BI, or longer units compared with UNI, is not only related to the restraint information, but also to the size of the corpus.

Table 4. **Results of disambiguation**

Number / Ambiguous words	Total	Correct	
		Before WSD	After WSD
什么(What/some)	230	168	172
几(how many/a few)	108	91	90
多少(how many/a lot)	90	50	54
还是(or/is still)	30	17	20
有没有(if/might be)	29	9	16

Table 5. **Contribution of each topic(%)**

Topic	Accuracy	Topic	Accuracy
Baseline	78.11	<C-O>	78.10
<G>	78.11	<O-I>	77.05
<P>	**78.21**	<E>	78.06
<F>	**78.17**	<O>	78.08
<T>	78.08	TP9	78.29
<C-I>	**78.26**	**TP3**	**78.36**

Table 6. **Accuracies(%) of DA recognition in SWBD corpus**

Feature Combination	Accuracy(%)
UNI	59.08
UNI+DSB	59.46
UNI+BI	*59.21*
(UNI+BI)_PST	60.92
(UNI+BI)_WT	61.17
(UNI+BI)_WT_UL	61.68
(UNI+BI)_WT_UL_FQ100	62.98
UNI+POS	61.83
UNI+BNP	63.18
(UNI+BNP)_FQ100_UL	64.76
(UNI+POS+BNP+BI)_FQ100_WT _UL+DSB	**64.92**

In the SWBD corpus, we also get 5.84% improvement of accuracy using selected features and combination rules as shown in Table 6.

5. Conclusion

In this paper, we introduce several textual features and propose a novel leveled feature structure for DA recognition. Comparative experiments are carried out to find out valuable features and combinations. The method has been evaluated on both labeled Chinese human-human dialog corpus, and public corpus SWBD. The experimental results on both corpora show significant improvement using the selected features and feature combination rules.

Especially, several effective features are first utilized in DA classification, including BNP in sentence structure and AP in discourse structure. Constraint information proposed in this paper is also remarkable for making better use of familiar features.

In future work, to further improve DA recognition accuracy, we will study the AP structure and make use of longer units. In addition, the application of DA, such as in spoken language translation and spoken dialog system will be considered.

6. Acknowledgments

The authors would like to thank Dr. Hwee Tou Ng for his beneficial comments and suggestions. The research work described in this paper has been supported by the Natural Science Foundation of China under grants 60723005, 90820303 and also supported by the Hi-Tech Research and Development Program (863) of China under grant 2006AA01Z194.

7. References

[1] J. Ang, Y. Liu, and E. Shriberg. 2005. Automatic Dialog-act Segmentation and Classification in Multiparty Meetings. In *Proceedings of the 30th ICASSP*, Philadelphia.

[2] J. L. Austin. 1962. How to do Things with Words. *Clarendon Press, Oxford*.

[3] S. Bangalore, G. D. Fabbrizio, and A. Stent. 2006. Learning the Structure of Task-driven Human-Human Dialogs. In *Proceedings of ACL 2006*. Sydney, July 2006. Pages 201-208.

[4] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg. 2004. Meeting Recorder Project: Dialog-act Labeling Guide. ICSI Technical Report TR-04-002. International Computer Science Insitute.

[5] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. In *Proceedings of ACL 2004*. Pages 669-676.

[6] J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of ICASSP 1992*. Volume 1, Pages 517-520.

[7] B. J. Grosz, and C. L. Sidner. 1986. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12(3):175-204.

[8] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. wooters. 2003. The ICSI Meeting Corpus. In *Proceedings of ICASSP 2003*, Hong Kong.

[9] G. Ji, and J. Bilmes. 2005. DA Tagging Using Graphical Model. In *Proceedings of ICASSP 2005*, Philadelphia.

[10] D. Jurafsky, L. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL Labeling Project Coder's Manual, Draft 13. Technical Report 97-02, University of Colorado Institute of Cognitive Science.

[11] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl. 1998. Lexical, Prosodic, and Syntactic Cues for Dialog Acts. In *Discourse Relations and Discourse Markers: Proceedings of the Conference*, Pages:114-120.

[12] P. Kral, C. Cerisara, and J. Kleckova. 2006. Automatic Dialog Acts Recognition Based on Sentence Structure. In *Proceedings of ICASSP 2006*, Toulouse, France. Pages 61-64.

[13] S. C. Levinson. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

[14] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T., E. Shriberg, and A. Stolcke. 2001. The Meeting Project at ICSI. *Human Language Technologies Conference*, San Diego.

[15] N. Reithinger, and E. Maier. 1995. Utilizing Statistical Dialog Act Processing in Verbmobil. In *Proceedings of ACL 1995*. MIT, Cambredge, MA. Pages 116-121.

[16] V. K. R. Sridhar, S. Narayanan, and S. Bangalore. 2008. Enriching Spoken Language Translation with Dialog Acts. In *Proceedings of ACL 2008, Short Papers(Companion Volume)*. Columbus, Ohio, USA, June 2008. Pages 225-228.

[17] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer. 2000. Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339-373.

[18] D. Surendran, and G.-A. Levow. 2006. DA Tagging with Support Vector Machines and Hidden Markov Models. In *Proceedings of Interspeech*, Pittsburgh, PA.

[19] D. Verbree, R. Rienks, and D. Heylen. 2006. Dialog-Act Tagging Using Smart Feature Selection; Results on Multiple Corpora. In *the first International IEEE Workshop on SLT*, Palm Beach, Aruba.

[20] M. Walker, and R. Passonneau. 2001. DATE: A Dialog Act Tagging Scheme for Evaluation of Spoken Dialog Systems. In *Proceedings of HLT 2001*, San Diego.

[21] F. Xu, C. Zong, and J. Zhao. 2006. A Hybrid Approach to Chinese Base Noun Phrase Chunking. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney. Pages 87-93.

[22] K. Zhou, C. Zong, H. Wu, and H. Wang. 2008. Predicting and Tagging DA with SVM and MDP. In *Proceedings of ISCSLP 2008*. December 16-19, 2008. Kunming, China. Pages 293-296.