What information is helpful for dependency based Semantic Role Labeling

Yanyan Luo **Kevin Duh** **Yuji Matsumoto**
Computational Linguistics, Nara Institute of Science and Technology
Takayama, Ikoma, Nara 630-0192, Japan
{yanyan-l; kevinduh; matsu}@is.naist.jp

Abstract

Semantic Role Labeling (SRL) is an important task since it benefits a wide range of natural language processing applications. Given a sentence, the task of SRL is to identify arguments for a predicate (target verb or noun) and assign semantically meaningful labels to them. Dependency parsing based methods have achieved much success in SRL. However, due to errors in dependency parsing, there remains a large performance gap between SRL based on oracle parses and SRL based on automatic parses in practice. In light of this, this paper investigates what additional information is necessary to close this gap. Is it worthwhile to introduce additional dependency information in the form of N-best parse features, or is it better to incorporate orthogonal non-dependency information (base chunk constituents)? We compare the above features in a SRL system that achieves state-of-the-art results on the CoNLL 2009 Chinese task corpus. Our findings suggest that orthogonal information in the form of constituents is much more helpful in improving dependency based SRL in practice.

1 Introduction

In recent years, SRL has become an important component in many kinds of deep natural language processing applications, such as question answering (Narayanan and Harabagiu, 2004), event extraction (Riedel and McCallum, 2011), document categorization (Persson et al., 2009). SRL aims at identifying the semantic relations between predicates in a sentence and their associated arguments, with these relations drawn from a pre-specific list of possible semantic roles for corresponding predicates. Syntax information is essential in SRL systems. To date, both constituent parsing and dependency parsing based SRL have been investigated (Xue, 2008; Johansson and Nugues, 2008), with dependency based systems giving superior results in CoNLL 2008 (Surdeanu et al., 2008) and CoNLL 2009 shared tasks (Hajič et al., 2009).

However, the performance gap is still quite large between SRL systems using oracle "perfect" dependency parses and SRL systems using automatic dependency parses. We observe as much as 10% F-score difference in our experiments. Clearly, errors in the 1-best dependency parse affects SRL prediction. This leaves an open question: in order to improve dependency based SRL, is it more worthwhile to incorporate more dependency information (in the form of N-best parse), or to incorporate an entirely separate source of information, such as base phrase chunks? We perform such an analysis in this paper, using a state-of-the-art Chinese SRL system.

Our findings suggest that constituent information such as chunking nicely complements dependency based SRL, achieving more improvements compared to N-best dependency information. Finally, we also report the best results to date on the CoNLL 2009 Chinese shared task.

2 Related Work

The bulk of previous work on automatic SRL has primarily focused on using full constituent parse of sentences to define argument boundaries and to extract relevant information for training classifiers. However, there have been some attempts at relaxing the necessity of using syntactic information derived from full parse trees. Sun et. al (2009) and Hacioglu et. al (2004) addressed the SRL problem on the basis of shallow syntactic information at the level of phrase chunks. In their approach, SRL is formulated as a sequence label-

International Joint Conference on Natural Language Processing, pages 781–787,
Nagoya, Japan, 14-18 October 2013.

ing problem, performing IOB2 decisions on the syntactic chunks of a sentence. However, this method ignores the full syntactic parsing information entirely, and we believe that even the accuracy of full syntactic parsing is not ideal, it is still helpful for SRL. Moreover, their method is inapplicable to dependency based SRL since a chunk usually consists of successive words.

A substantial amount of research has focused on dependency-based SRL (Meza-Ruiz and Riedel, 2009; Luo et al., 2012) since the CoNLL-2009 shared task and rich linguistic features (Zhao et al., 2009) are applied. For dependency related features, most studies focused on extracting them from the best dependency result. Johansson and Nugues (2008) tried to use N-best dependency parsing results. In their work, they applied 16-best dependency trees to generate predicate-argument structures and applied both syntactic trees and predicate-argument structures to a linear model. This model reranks the predicate-argument structures and the top 16 dependency trees at the same time. Though their work suggests that N-best dependency parsing can enhance the SRL, little is known about how the N-best dependency parsing related features perform on SRL.

3 Dependency based SRL Model

First, we define an instance as a predicate word and its corresponding argument words. If there are m predicates in a sentence, then there will be m instances. Given an instance $X = \{x_1, \ldots, x_p, \ldots, x_n\}$ with the predicate position p, we want to find the corresponding sequence of argument labels and predicate sense $S = a_1, a_{p-1}, P, a_{p+1}, \ldots, a_n = \langle P, A \rangle$. Each a_i for the i-th word in the instance X is drawn from a set of tags $T(A)$ which contains all the semantic role labels in the corpus and which follows the definition criteria in Chinese PropBank. In addition, the special label $NONE$ is added to $T(A)$. Words, labeled as $NONE$, are not arguments for the predicate. As for P, this is a member of a sense set $T(x_p)$ which contains all possible senses of predicate word x_p. We propose two sorts of label assignment models Pr_{local} and Pr_{global}. The former can incorporate local features only; the latter can incorporate also global features. We use three types of local feature sets: F_P, F_A, F_{PA} and one global feature set F_G. These type definitions are the same as those in Watanabe et. al (2010).

3.1 Predicate Sense Disambiguation and SRL with a Local Model

Since the predicate cannot be an argument of itself for Chinese, we define the following local probabilistic model for argument classification and predicate sense disambiguation.

$$Pr_{local}(S|X) = \prod_{i=1(i\neq p)}^{n} Pr(a_i|P, X, i, p) \cdot Pr(P|X, p) \quad (1)$$

where $Pr(a_i|P, X, i, p)$ and $Pr(P|X, p)$ are estimated according to the following equation:

$$Pr(a_i|P, X, i, p) = \frac{1}{Z^A(X)} \exp\{ \sum_{f_{A_j} \in F_A} \lambda_{f_{A_j}} f_{A_j}(a_i, X) + \sum_{f_{PA_k} \in F_{PA}} \lambda_{f_{PA_k}} f_{PA_k}(a_i, X, p, P)\},$$

$$Pr(P|X, p) = \frac{1}{Z^P(X)} \exp\{ \sum_{f_{P_l} \in F_P} \lambda_{f_{P_l}} f_{P_l}(X, p, P)\},$$

where Z^A and Z^P are normalization functions, i.e.,

$$Z^A = \sum_{a_i \in T(A)} \exp\{ \sum_{f_{A_j} \in F_A} \lambda_{f_{A_j}} f_{A_j}(a_i, X) + \sum_{f_{PA_k} \in F_{PA}} \lambda_{f_{PA_k}} f_{PA_k}(a_i, X, p, P)\};$$

$$Z^P = \sum_{P \in T(x_p)} \exp\{ \sum_{f_{P_l} \in F_P} \lambda_{f_{P_l}} f_{P_l}(X, p, P)\};$$

f are the features with associated weight λ learned via training.

3.2 Predicate Sense Disambiguation and SRL with the Global Model

Global information is known to be useful in SRL (Nakagawa, 2007). We propose a global probabilistic model Pr_{global} here for SRL as follows:

$$Pr_{global}(S|X) = \frac{1}{Z} Pr_{local}(S|X) \cdot exp \left\{ \sum_{f_{G_m} \in F_G} \lambda_{f_{G_m}} f_{G_m}(S, X) \right\} \quad (2)$$

where Z is a normalizing factor over all candidate sequences $S(X, p)$ (set of possible configurations of semantic tags and predicate senses given X and predicate location p). To get the whole sequence of S, we need to perform a computationally expensive search. As done in previous work (Watanabe et al., 2010), we use a simple approach,

Type	%Error	#Error/#Occurrence
C	49.4%	7,162/14,497
G	88.62%	109/123
O	80.71%	3,175/3,934

Table 1: The distribution of SRL errors on development corpus by the joint model.

n-best relaxation. Unlike the $Pr_{local}(S|X)$, the product of probability distributions of each word, the probability distribution $Pr_{global}(S|X)$ is calculated by feature functions f_G defined on an instance X with assignment S. Thereby, we can use any information in an instance without the independence assumption for assignments of words in it.

3.3 Error Analysis for Dependency-based SRL

Using the gold parse of dependency relations between a predicate and its arguments and according to these relations, we classified SRL errors into following three types.

- **C**: children of a predicate should be arguments but they are tagged incorrectly.
- **G**: grand children of a predicate should be arguments but they are tagged incorrectly.
- **O**: others

Table 1 shows the distribution of three errors observed in the development corpus after tagging by our joint model. For example, there are a total of 14,497 arguments that are children of predicates and among them, and 7,162(49.4%) are errors.

4 Results and Discussion

4.1 Experimental Setting

We used the Chinese dataset provided by CoNLL-2009 shared task for experiments. For comparison, two kinds of dependency parsing results are provided, the first is from MALT parser, the second is from second-order MST parser.

As for chunking information, we used the chunk definition presented in (Chen et al., 2006) to extract chunks from Chinese Tree Bank as training corpus. The line CH in Figure 1 shows the definition of chunks. In this example, " 金融工作"(finance work) is a noun phrase and is composed by two nouns.

With the Inside/Outside representation for proper chunks and the following feature templates, where x_0 is the current word, a CRF++[1] is trained for Chinese chunking task.

- Uni-gram word/POS tag features: x_{-2}, x_{-1}, x_0, x_{+1} and x_{+2}.
- Bi-gram word/POS tag features: $x_{-2}x_{-1}$, $x_{-1}x_0$, x_0x_{+1} and $x_{+1}x_{+2}$.

4.2 Features

Most of features templates are "standard" which have been widely used in previous dependency-based SRL research (Johansson and Nugues, 2008; Luo et al., 2012). We do not explain "standard" features, however, we give a detailed description of the features used in this work.

4.2.1 Base Phrase Chunking Related Features

In Figure 1, obviously, words in chunks do not have equal importance for SRL. Headwords represent the main meaning of the chunks. The base phrase chunking related features shown in Table 2 are only applied to these headwords. For other words in chunks, only lemma and POS information is used. The rules described in Sun and Jurafsky (2004) are used to extract headwords. Verb class in Table 2 is represented similarly as $Verb.C1C2$, which means this $verb$ has two senses. For its first sense, it has one core argument and for its second sense, it has two core arguments. These verb classes are extracted from Chinese PropBank (Xue, 2008).

4.2.2 Features from N-best Dependency Parsing

According to the statistics of development corpus, it is found that about 78.13% arguments are children of predicates. Even if its error percentage shown in Table 1 is less than 10%, the total error number is also considerable. If we can reduce the head errors for dependents, the C errors caused by dependency parsing errors should be decreased, and SRL tagging results would be improved. Under this hypothesis, we simply extracted the following features from every parse tree in the N-best list which are generated using second-order MST parser. These features are also included in the "standard" feature set when $N = 1$.

[1] http://crfpp.sourceforge.net/

WORD	去年	西藏	金融	工作	取得	显著	成绩
POS	NN	NR	NN	NN	VV	JJ	NN
CH	[NP]	[NP]	[NP	]	[VP]	[ADJP]	[NP]
TAG	B-NP	B-NP	B-NP	I-NP	B-VP	B-ADJP	B-NP
SRL	TMP	NONE	NONE	A0	取得.01	NONE	A1

Figure 1: Chunking information for a predicate-argument structure.

Feature Name	Description
Chunk features	chunk tag of headword with IB representation (e.g. $B-NP$) chunk tag of the chunk where the headword belongs to the number of words in a chunk the POS sequence of words in a chunk, for example, "金融工作" (finance work) is "NN_ NN" the position of the chunk with respect to the predicate(Position). There are three possible values: "before", "after" and "here". the conjunctions of Position and headword, predicate and verb class the conjunctions of Position and POS of headword, predicate and verb class lemma/POS of one word immediately before/after of the chunk
Path features	a chain of chunk types between the headword and the predicate. the length of the chunk chain between the headword and the predicate For example, chain of chunk types between headword "工作" and predicate "取得" is "NP-VP" and the length of the chunk chain is 2.

Table 2: Chunking related feature template for experiments.

$Arguments'heads$: lemma/pos; lemma and pos; dependency label; whether they are predicates.

$Position$: position of the argument candidates with respect to the predicate positions in the tree; position of the heads of the argument candidates with respect to the predicate position in the sentence.

$Chain$: the left-to-right chain of the dependency labels of the predicate's dependents.

4.3 SRL Performance

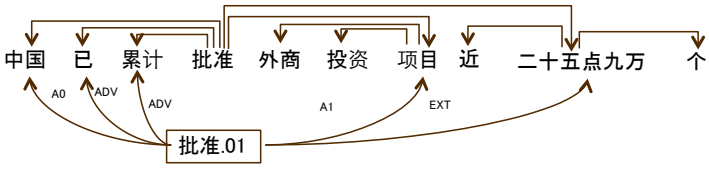
The overall performance of SRL is calculated using the semantic evaluation metric of the CoNLL-2009 shared task scorer[2]. Table 3 gives the comparison of SRL performance before and after adding the proposed base phrase chunking related features on the test data. Lines with $-/+$ show the SRL performance without/with base phrase chunking related features. As seen in this table, without gold dependency parse, the best SRL is up to 80.52 in F_1 score. To the best of author's knowledge, there are few Chinese SRL results more than 80%.

[2] http://ufal.mff.cuni.cz/conll2009-st/eval09.pl

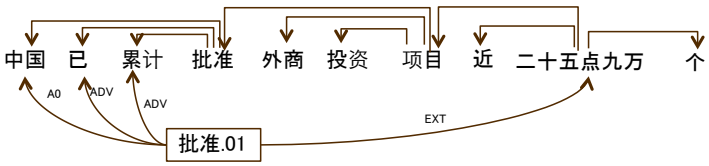
	P(%)	R(%)	F_1(%)
Gold parsing $-$	88.68	86.30	87.47
Gold parsing $+$	**90.03**	**87.71**	**88.86**
MALT $-$	82.64	72.68	77.34
MALT $+$	**84.17**	**74.67**	**79.13**
MST-2 $-$	83.01	75.39	79.02
MST-2 $+$	**84.49**	**76.92**	**80.52**

Table 3: SRL results without/with base phrase chunking information.

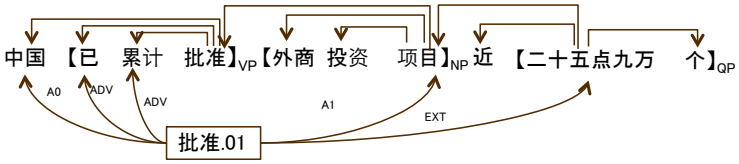
Although comparing the lines with $-$, it shows dependency parsing play the central role in Chinese SRL as expected. Comparing their corresponding lines with $+$, Chinese SRL can still benefit a lot from shallow parsing information. An example from the corpus is shown in Figure 2. Figure 2a shows the gold dependency parsing result and the gold predicate argument structure; Figure 2b shows the dependency parsing result from MALT parser and the predicate argument structure as a result of the predicted parse; Figure 2c shows the predicate argument structure which is predicted after adding base phrase chunking re-



(a) Gold SRL result with gold dependency tree



(b) Wrong SRL result brought by MALT parse errors



China has approved nearly 259,000 foreign investment projects.

(c) SRL result corrected by introducing base phrase chunking related features

Figure 2: An example that the argument prediction error brought by MALT parse errors is corrected by introducing base phrase chunking related features.

	N-best	P(%)	R (%)	F_1(%)
MST-2−	1	**83.01**	75.39	79.02
MST-2−	3	82.52	**77.16**	79.75
MST-2−	5	82.74	77.10	**79.82**
MST-2−	10	82.44	76.98	79.62

Table 4: SRL results with N-best dependency parsing related features.

N-best	Correct (#)	Error(#)	Noise(#)
1	18,428	3,176	-
3	19,071	2,533	4,636
5	19,392	2,212	5,667
10	19,738	1,866	7,699

Table 5: Dependency accuracy and the noise changes with different N.

lated features. In Figure 2c, the subscripts stand for chunk types. From Figure 2b, it can be seen that the argument $A1$ is not identified by the dependency based SRL because of dependency errors. Comparing Figure 2b and 2c, we can see that after adding the base phrase chunking related features, this SRL error brought by dependency parsing errors is corrected.

Line MALT+ and line MST-2− show that even the dependency parsing result from MALT is not better than that from second order MST, with the aid of chunking related features, Chinese SRL can still get comparable results.

Table 4 shows the Chinese SRL results after adding the N-best dependency parsing related features. It is not surprising that SRL can get better performance when $N > 1$, because the larger N, a more accurate dependency parsing results can be likely obtained. When $N = 5$, SRL gets the best performance 79.82 in F_1 with 0.8 point improvement.

However, the improvement declines when $N = 10$. A larger N may result in adding more accurate dependency parsing, however, it can also result in including more noises. For the MST parser using second order algorithm, Table 5 shows how the choice of the value of N affects the dependency parsing. The Correct(#) column represents the number of cases where the correct parent of an argument is predicted within the N-best. For example, in 3-best, it counts the number of arguments where their parents are correctly predicted in at least one of the 3 predictions. In the case where the parent is not predicted in any tree, they are counted as an error, as listed in the second column. The third column (Noise), is defined under

	N-best	P(%)	R (%)	F_1(%)
[Björkelund, 2009]	-	82.42	75.12	78.60
[Meza-Ruiz, 2009]	-	82.66	73.36	77.73
[Zhao, 2009]	-	80.42	75.20	77.72
MST-2 +	1	**84.49**	76.92	80.52
MST-2 +	3	83.81	**78.51**	**81.07**
MST-2 +	5	83.71	78.40	80.97

Table 6: SRL results with base phrase chunking information and N-best parsing related features.

	N-best	C(%)	G(%)	O(%)
MST-2−	1	25.37	86.93	59.06
MST-2−	3	22.83	78.43	56.84
MST-2−	5	22.83	78.43	57.36
MST-2+	1	23.93	86.93	54.70
MST-2+	3	21.5	76.47	53.05
MST-2+	5	21.66	76.47	53.38

Table 7: SRL error changes with different features

a hypothesis: correct dependency relations generate correct SRL results, wrong dependency relations generate incorrect SRL results. It represents the number of wrong dependency relations in Correct case which can cause bad influence on SRL results. For example, if 3 best heads for an argument are top-1, top-2, top-3 respectively, and top-1 is the correct one, then this case is a Correct case and the number of noise are 2; if none of the three results are correct, then this case is an Error case, and no noise. From this table, it obviously indicates that the benefit for dependency parsing brought by a larger N is less than the noise brought by the N.

With Tables 3 and 4, it can be seen that SRL benefits more from chunking related features than from N-best parse related features.

Table 6 shows the the results of Chinese SRL after adding base phrase chunking information and N-best parsing related features and gives the comparison with the previous work. From Tables 4 and 6 we can see that after adding the chunking related features, the impact of N-best parsing related features is a little reduced.

4.4 Discussion

In Section 4.3, we see that both chunking and N-best parsing related features are helpful for Chinese SRL to some extent. In order to understand how they affect SRL, we analyze the results from three types of errors introduced in Section 3.3. Table 7 shows the error changes when different features are added.

Since accurate dependency information is not always available, the three types of errors should become larger when automatic dependency parsers are used. From Tables 1 and 7, the C and O errors increased as expected, while G decreased. The main reason is that arguments, that are grandchildren of predicates, are relocated in the dependency trees because of dependency errors, and these locations make them easier to be tagged. From the first and fourth rows, they suggest that shallow parsing information are helpful to reduce the C and O errors. Comparing the fourth line with second and third rows, they explain why SRL achieves more improvements from chunking than from N-best dependency. When N changed from 1 to 3, the errors decreased obviously, however, when the $N = 5$, there are no obviously different changes.

5 Conclusions and Future Work

In this paper, we introduce additional information: base phrase chunking and N-best dependency parsing related features to a dependency based SRL system and investigate the benefit that our Chinese SRL model can get from them. Evaluations on the CoNLL 2009 Chinese corpus show that chunking information well complements dependency based SRL, achieving more improvements compared to N-best dependency information. With those additional features, our dependency based SRL achieves the best result on the same Chinese corpus to our knowledge. Furthermore, while all our experiments are for Chinese, it is possible to design experiments for other languages with our models.

Our experiment results show that we are not limited to increase SRL performance via more accurate syntactic parsing, but that we can explore other information, which is easier to get and is helpful for SRL. This also guides our future work. In our future work, we would like to explore more features and their influence on SRL.

References

Anders Björkelund, Love Hafdell and Pierre Nugues. 2009. Multilingual Semantic Role Labeling. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning(CoNLL)*, pp. 43-48.

Wenliang Chen, Yujie Zhang and Hitoshi Isahara. 2006. An Empirical Study of Chinese Chunking. *Proceedings of the COLING/ACL on Main conference poster sessions* , pp. 97-104.

Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James H.Martin and Daniel Jurafsky. 2004. Semantic Role Labeling by Tagging Syntactic Chunks. *Proceedings of the 8th Conference on CoNLL-2004, Shared Task* .

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antòia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* , pp. 1-18.

Richard Johansson and Pierre Nugues. 2008. Dependency-based Semantic Role Labeling of PropBank. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 69-78.

Luo Yanyan, Asahara Masayuki and Matsumoto Yuji. 2012. Robust Integrated Models for Chinese Predicate-Argument Structure Analysis. *China Communications* , 9(3): pp. 10-18.

Ryan McDonald, Koby Crammer and Fernando Pereira. 2005. Online Large-margin Training of Dependency Parsers. *Proceeding of the 43th Annual Meeting on Association for Computational Linguistics*, pp.91-98.

Ivan Meza-Ruiz and Sebastian Riedel. 2009. Jointly Identifying Predicates, Arguments and Senses Using Markov Logic. *Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics(HLT/NAACL-2009)*, pp. 155-163.

Tetsuji Nakagawa. 2007. Multilingual Dependency Parsing Using Global Features. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007* , 34(2): pp. 952-956.

Srini Narayanan and Sanda Harabagiu. 2004. Question Answering Based on Semantic Structures. *Proceeding of the 20th International Conference on Computer Linguistics*, pp. 693-701.

Jacob Persson, Richard Johansson and Pierre Nugues. 2009. Fast and Robust Joint Models for Biomedical Event Extraction. *NODALIDA 2009 Conference Proceedings* , pp.142-149.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text Chunking using Transformation-Based Learning. *Proceedings of the 3rd Workshop on Very Large Corpora* , pp. 88-94.

Sebastian Riedel and Andrew McCallum. 2011. Fast and Robust Joint Models for Biomedical Event Extraction. *Proceeding of the 2011 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Languages Learning* , pp. 1-12.

Honglin Sun and Daniel Jurafsky. 2004. Shallow Semantic Parsing of Chinese. *Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics(HLT/NAACL-2004)* , pp. 1249-256.

Weiwei Sun, Zhifang Sui, Meng Wang and Xin Wang. 2009. Chinese Semantic Role Labeling with Shallow Parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* , pp. 1475-1483.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez and Joakim Nivre. 2008. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic dependencies. *Proceedings of the 12th Conference on Computational Natural Language Learning* , pp. 157-177.

Kristina Toutanova, Aria Haghighi and Christoper D. Manning. 2008. A Global Joint Model for Semantic Role Labeling. *Computational Linguistics* , 34(2): pp.161-191.

Yotaro Watanabe, Masayuki Asahara and Yuji Matsumoto. 2010. A Structured Model for Joint Learning of Argument Roles and Predicate Senses. *PProceedings of the ACL 2010 Conference Short Papers* , pp. 98-102.

Nianwen Xue. 2008. Labeling Chinese Predicates with Semantic Roles. *Computational Linguistics* , 34(2): pp. 225-255.

Hai Zhao, Wenliang Chen, Chunyu Kit and Guodong Zhou. 2009. Multilingual Dependency Learning:A Huge Feature Engineering Method to Semantic Dependency Parsing. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task* , pp. 55-60.