Can Selectional Preferences Help Automatic Semantic Role Labeling?

Shumin Wu
Department of Computer Science
University of Colorado Boulder
shumin@colorado.edu

Martha Palmer
Department of Linguistics
University of Colorado Boulder
mpalmer@colorado.edu

Abstract

We describe a topic model based approach for selectional preference. Using the topic features generated by an LDA model on the extracted predicate-arguments over the Chinese Gigaword corpus, we show improvement to our state-of-the-art Chinese SRL system by 2.34 F1 points on arguments of nominal predicates, 0.40 F1 point on arguments of verb predicates, and 0.66 F1 point overall. Moreover, similar gains were achieved on out-of-genre test data, as well as on English SRL using the same technique.

1 Introduction

It's long been theorized that selectional preferences (SP)/semantic constraints can improve automatic semantic role labeling (SRL). And while there have been several publications showing positive effects of SP, the evaluations have been dominated by pseudo-disambiguation. Zapirain et al. (2013) demonstrated end-to-end SRL improvement on arguments of English verb predicates by using a combination of lexical resources and distributional similarity based SP. However, the margin of improvement is a modest 0.4 F1 point (on WSJ) over a baseline system with performance over 4 F1 points lower than the top system in CoNLL-2005 (Carreras and Màrquez, 2005). These results may not be convincing enough to motivate the incorporation of SP when building an SRL system. One reason for the small improvement may be that arguments of a verb predicate are highly constrained by the underlying syntactic parse, and SP features that could disambiguate between role types are often negated by parse errors. With the recent extension of PropBank SRL to nominal and adjective predicates, preposition relationships, light-verb constructions, and abstract meaning representation (Bonial et al., 2014; Banarescu et al., 2013), it may be time to revisit SP for SRL. We hypothesize that SP will provide a greater benefit to nominal SRL, especially on a language with lower parsing accuracy.

In this paper, we apply SP to Chinese SRL (which has few morphological clues that impacts parsing accuracy) for arguments of both verb and nominal predicates using Chinese Gigaword. Our hypothesis, that SP will provide a greater benefit for nominal predicates than for verbal predicates, is verified by our results. We achieve a 2.34 F1 point improvement to our Chinese SRL system on arguments of nominal predicates, 0.40 F1 point on arguments of verb predicates, and 0.66 F1 point overall.

2 Previous Work on Selectional Preference

Inducing selectional preferences from corpus data was first proposed by Resnik (1997) for sense disambiguation. He generalized seen words using the WordNet (Fellbaum, 1998) hierarchy. Gildea and Jurafsky (2002) applied SP to automatic SRL by clustering extracted verb-direct object pairs, resulting in modest improvements. This syntactic signature based selectional preference technique has also been successfully extended and applied to unsupervised SRL by Lang and Lapata (2011) (using split-merge role clustering), as well as Titov and Klementiev (2012) (using a distance-dependent Chinese Restaurant Process prior for role clustering). Zapirain et al. (2013) improved the end-to-end perfor-

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 222–227,
Denver, Colorado, June 4–5, 2015.

mance of an English PropBank SRL system by 0.4 F1 points using a variety of word similarity measures, from WordNet hierarchy distance to distributional similarity measures.

Ritter and Etzioni (2010) reasoned that the set of hidden variables modeled by latent Dirichlet allocation (LDA) naturally represents the semantic structure of a document collection, and the topics generated can be viewed as the latent set of classes that store preferences. The work utilizes LinkLDA, a variant of the standard LDA that models two sets of distributions for each topic simultaneously, with the resulting topics encoding the mutual constraints of a pair of arguments for the same predicate. Séaghdha and Korhonen (2014) also proposed SP w/ the LDA variants ROOTH-LDA and LEX-LDA.

There has also been work on Chinese selectional preferences, both lexical resource (HowNet) based and corpus based (Jia et al., 2011; Jia et al., 2013). The authors found the LDA corpus based SP improved over the HowNet based SP on pseudo-disambiguation. All of these results encouraged us to also attempt an LDA based approach to SP.

3 Selectional Preference for SRL

3.1 SP Representation

Some of the most discriminative SP models used by Zapirain et al. (2013) relied on distributional similarity computed over dependency relationships (provided by Lin (1998)). For example, in "*John lent Mary the book.*", we would extract *John-nsubj*, *Mary-iobj*, *book-dobj* for the predicate *lend*. While this has proven to be of higher quality than pure word co-occurrence based similarity, it may not be optimal for semantic-based processing. With nominal SRL, a large portion of the arguments (around 50% in Chinese PropBank) are not the direct syntactic dependents of the predicate: in figure 1, because of a light verb-like construction, all the arguments of 欢迎/*welcome* are the syntactic dependents of 表示/*express*. To address this, we directly extract SP of the predicates by running our SRL system over the unannotated corpus. For our example, we would extract *John-Arg0*, *Mary-Arg2*, *book-Arg1* for *lend*.

3.2 SP with LDA-based Topic Model

Our approach to modeling selectional preferences (SP) follows a relatively straightforward application of LDA to a set of predicate-argument instances derived from a corpus. In the standard LDA model, a document d is represented by a bag of words and is drawn from a multi-nominal Dirichlet θ_d over topics. The resulting model is a probability distribution of each word amongst the topics.

For the SRL application, we treat each extracted argument (represented by the $(label, headword)$ pair) as a "word", and the collection of arguments for all instances of a particular predicate as a "document". The generated topics would then contain arguments sharing a similar set of predicates. With this definition, we allow different role labels to share the same topic (though it does not encode role constraints quite like LinkLDA, ROOTH-LDA, etc). For prepositional phrases, we used the dependent of the preposition as the head word since the preposition can often be omitted in Chinese.

3.3 SRL Filtering

Building selectional preferences by means of using the output of an SRL system is unlikely to improve the same SRL system unless one filters out the lower quality labels (in earlier experiments where we performed no filtering, this was indeed the case). We ran SRL on the unannotated corpus using a logistic regression model and filtered out the low probability output. To balance between precision and recall, we set a hard 0.5 probability cutoff and discounted the occurrences of the rest using the label probability.

Since we can extract higher quality SP from the output of a better performing SRL system, we can iteratively improve our SRL system by re-extracting SP using a retrained (SP enhanced) SRL system. We arrived at diminishing returns after one additional iteration (of training SRL, extracting SP, and retraining SRL w/ new SP).

4 SRL Implementation

Our Chinese SRL system follows the standard (English) approach where the SRL task is posed as a multi-class classification problem requiring the identification of argument candidates for each predicate and their argument types using a set of lexical

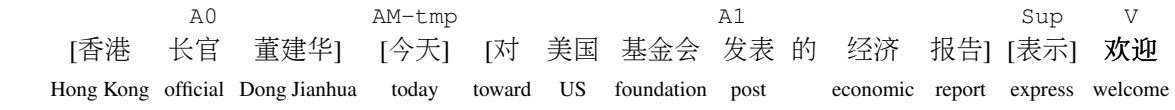

Figure 1: Chinese nominal predicate translated to English verb predicate

and syntactic features (predicate word, constituent head, path, syntactic frame, etc). While the top SRL systems from CoNLL-2005[1] and some subsequent systems use multiple parses for structural inference, we instead implement a 2-stage argument label classification system on a single input parse: the argument set found by the first classifier is used as an additional feature for the second classifier (to identify missing or duplicate argument label types).

4.1 Selectional Preference

The LDA topic model produces a probability distribution of words (represented here by the ($label$, $headword$) pair) over topics. For the SRL task, argument candidates with topic distributions similar to those of the arguments found in the training set are likely to be permissible. Ideally, we would use these distributions directly. Since our SRL system was designed to accept lexical (binary) features only (for training/decoding performance), we pared the distribution down to at most 3 topics for each $label$ type and excluded words that do not have high affinity to a few topics (sum of the probability of the top 3 topics $< 50\%$) to prevent diluting the discriminative power of the topic feature. We used the resulting list of ($label$, $topic\_id$) pairs for each word as the selectional preference feature for each encountered constituent in the Chinese SRL system.

During the normal LDA inference stage, using the learned topic model, a predicate instance ("document") will be assigned a probability distribution over topics based on its arguments, and each argument will be assigned a specific topic (or topic distribution). This could further constrain an argument's selectional preference within the context of the predicate instance and other arguments. For our system, we experimented with performing inference on the argument label set extracted from the first stage classifier and using the constrained argument topic distribution for the second stage classifier. However, we observed no improvement, likely because there are only a few arguments for each predicate instance.

[1]We use CoNLL-2005 instead of CoNLL-2009 for comparison because our SRL system is based on constituent parses.

5 Experiment

5.1 Setup

Our Chinese SRL system is trained on Chinese TreeBank 5.1 and Chinese PropBank 1.0. We used the standard: sections 81-885 for training, sections 41-80 for development, and sections 1-40, 900-931 for testing. We generated the training parses (with 10 fold cross-validation) and the test parses using the Berkeley parser[2] (5 split-merge cycles). The parser F1 score on the test sections is 82.73 as measured by ParseEval (Black et al., 1991).

We prepared the Chinese Gigaword[3] corpus with the Stanford Chinese Word Segmenter[4]. We performed LDA topic modeling using PLDA+ (Liu et al., 2011) and the recommended $\alpha = 50/topic\_cnt$, $\beta = 0.01$ values. We chose 2000 topics (tuned on the SRL performance of the development set rather than any topic based metrics). Table 1 lists some of the found topics (with the most frequent, relatively interesting, and least frequent headword, label pairs) using Chinese Gigaword.

5.2 Performance

As table 2 shows, the addition of the *SP* feature improved nominal SRL by 2.34 F1 points. Verb SRL improved by 0.40 F1 point and overall SRL improved by 0.66 F1 point. These F1 differences were all found to be statistically significant[5] ($p \leq 0.05$).

We also tested the system on Sinorama magazine and other out-of-genre sections (broadcast conversation, broadcast news, web blog) in Chinese Prop-

[2]code.google.com/p/berkeleyparser/

[3]LDC2011T13

[4]nlp.stanford.edu/software/segmenter.shtml

[5]$SIGF$ (www.nlpado.de/%7esebastian/software/sigf.shtml), using stratified approximate randomization test (Yeh, 2000)

topic	headword:argument_label pairs
emergency response	破坏/damage:Arg1 阻止/stop:Arg1 制造/fabricate:Arg1 寻找/search:Arg1 自杀/suicide:Arg1 ... 灭火/extinguish:Arg1 敲诈/blackmail:Arg1 挣脱/break_free:Arg1 东山再起/comeback:Arg1
government agency	海关/custom:Arg0 联合会/union:Arg0 务部/work_department:Arg0 旅游局/travel_department:Arg0 统计局/census:Arg0 ... 部会/ministries:Arg0 边检站/checkpoint:Arg0 财政局/finance_bureau:Arg0
law & order	警方/police:Arg0 嫌犯/suspect:Arg1 男子/male:Arg1 到案/court_appearance:Arg1 公安/public_safety:Arg0 ... 巷/alley:Argm-loc 嘉义市/Chiayi_City:Argm-loc 哥伦比亚人/Columbian:Arg1
path	道路/road:Arg1 路/path:Arg1 大道/avenue:Arg1 ... 红地毯/red_carpet:Arg1 钢丝/steel_wire:Arg1 独木桥/plank_bridge:Arg1 ... 迷宫/maze:Arg1 侧门/side_entrance:Arg1 险棋/risky_move:Arg1
competition	比赛/competition:Arg1 决赛/final:Arg1 联赛/league_comp:Arg1 ... 考试/exam:Arg1 大选/election:Arg1 世乒赛/world_pingpong_match:Arg1 ... 加赛/playoff:Arg1 分团/sub-group:Arg0
moral & ethics	精神/spirit:Arg1 传统/tradition:Arg1 作风/style:Arg1 文明/civil:Arg1 ... 校风/school_spirit:Arg1 同舟共济/share_hard_time:Arg1 ... 幸福观/happy_outlook:Arg1 博爱/universal_love:Arg1

Table 1: Topics in Chinese Gigaword

system	nominal			verb	all
	p	r	f1	f1	f1
baseline	64.71	48.20	55.25	75.53	72.08
SP_{LDA}	65.70	**51.27**	**57.59**	**75.93**	**72.74**

Table 2: Chinese PropBank 1.0 results

sections	system	p	r	f1
Sinorama nominal	baseline	37.58	25.10	30.10
	SP_{LDA}	39.72	27.36	32.40
verb	baseline	67.13	50.37	57.55
	SP_{LDA}	67.56	50.59	57.86
4051-4411 (verb)	baseline	62.01	50.74	55.81
	SP_{LDA}	62.70	51.03	56.27

Table 3: Chinese PropBank 3.0 out-of-genre results

Bank 3.0. Only Sinorama has nominal SRL annotations. As table 3 shows, even though the absolute performance is much lower, SP improved the precision and recall in all cases, the nominal SRL score on Sinorama by 2.30 F1 points, and verb SRL score by 0.31-0.46 F1 point. Again, these F1 differences were statistically significant.

5.2.1 Comparison

Direct performance comparison with previous Chinese SRL systems is a bit difficult: Xue (2008), Zhuang and Zong (2010) trained the syntactic parsers with an additional 250K word broadcast news corpus found in Chinese TreeBank 6.0, while Sun (2010) only reported results using gold POS tags but no additional gold parses. However, as table 4 shows, for verb predicates, our system bests Xue's (2008) system by 4-7 F1 points with less parser training data and when tested with (but was not retrained to take full advantage of) gold POS tags besting Sun's (2010) system by 0.53 F1 point. For nominal predicates, our system bests Xue's (2008) system, by 1.9 F1 points on arguments of nominal predicates (since we have an integrated SRL system, the results are obtained by training both verb and nominal predicates, then using only the nominal classifier to classify the nominal predicates).

5.2.2 English SRL

We applied the same techniques to English SRL using the English Gigaword[7] corpus. We used 800 topics (w/ lemmatized headwords) tuning on the

[6]Verb results are from SRL systems trained on verbs only. Table 2 results are from SRL systems trained on all predicates.

[7]LDC2003T05

type	system	p	r	f1
verb	Xue 2008	76.8	62.5	68.9
	w/ gold POS	79.5	65.6	71.9
	Sun 2010 (gold POS)	81.03	**72.38**	76.46
	SP_{LDA}	82.74	70.96	76.40
	w/ gold POS	**82.81**	71.93	**76.99**
nominal	Xue 2008	62.9	53.1	57.6
	SP_{LDA}	**67.30**	**53.31**	**59.50**

Table 4: Chinese SRL comparison[6]

system	p	r	f1	$error_{\Delta}$
SwiRL	79.7	70.9	75.0	
Zapirain 2013	80.0	71.3	**75.4**	-1.60%
baseline	82.59	77.27	79.84	
SP_{LDA}	82.96	77.52	**80.15**	-1.54%

Table 5: English SRL comparison (CoNLL-2005 WSJ)

CoNLL-2005 development set. Compared to Zapirain et al. (2013) (table 5), our SP approach had a smaller (but still statistically significant) absolute F1 gain, with most of the gain coming from core argument type improvements. But with a much higher performing baseline system (one of the highest reported results using a single input parse per sentence), the error reduction rate is comparable.

6 Conclusion

We presented a LDA topic model based selectional preference approach to improving automatic SRL. Using SP extracted from a 63.6M sentence Chinese Gigaword corpus, we were able to improve on the results of an already competitive Chinese SRL system by 2.34 F1 points on nominal predicates, 0.40 F1 point on verb predicates, and 0.66 F1 point on the standard test set. More over, we obtained comparable improvement on out-of-genre data and demonstrated our technique is also applicable to English SRL. Given the margin of improvement on nominal SRL, which is not as well constrained by syntax as verb SRL, there are reasons to speculate the proposed technique could be applicable to other predicate type extensions of PropBank SRL.

As our first attempt at automatically deriving Chinese selectional preference, there is a lot of room for future improvement. Notably, these include techniques used for English SP such as computing similarity based on lexical resources (for Chinese - HowNet (Dong et al., 2010)), distributional similarity, latent word language model (Deschacht and Moens, 2009), different variants of LDA topic models, as well as taking advantages of argument constraints in parallel corpora to extract higher quality SP.

Acknowledgement

We gratefully acknowledge the support of the National Science Foundation CISE-IISRI-0910992, Richer Representations for Machine Translation, DARPA FA8750-09-C-0179 (via BBN) Machine Reading: Ontology Induction: Semlink+, and DARPA HR0011-11-C-0145 (via LDC) BOLT. This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider, 2013. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, chapter Abstract Meaning Representation for Sembanking, pages 178–186. Association for Computational Linguistics.

E. Black, S. Abney, S. Flickenger, C. Gdaniec, C. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. Procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 306–311, Stroudsburg, PA, USA. Association for Computational Linguistics.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. 2014. Propbank: Se-

mantics of new predicate types. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 152–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koen Deschacht and Marie-Francine Moens. 2009. Semi-supervised semantic role labeling using the latent words language model. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 21–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhendong Dong, Qiang Dong, and Changling Hao. 2010. Hownet and its computation of meaning. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 53–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Yuxiang Jia, Hongying Zan, and Ming Fan. 2011. Inducing chinese selectional preference based on hownet. In *Proceedings of the Seventh International Conference on Computational Intelligence and Security*, CIS2011, pages 1146–1149.

Yuxiang Jia, Hongying Zan, Ming Fan, , Shiwen Yu, and Zhimin Wang. 2013. Computational models for chinese selectional preferences induction. *International Journal of Advanced Intelligence*, 5(1):110–119, July.

Joel Lang and Mirella Lapata. 2011. Unsupervised semantic role induction via split-merge clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, June. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of ACL 1998*, ACL '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, and Maosong Sun. 2011. Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing. *ACM Trans. Intell. Syst. Technol.*, 2(3):26:1–26:18, May.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics*, pages 52–57, Washington, D.C. ACL.

Alan Ritter and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Diarmuid Séaghdha and Anna Korhonen. 2014. Probabilistic distributional semantics with latent variable models. *Computational Linguistics*, 40(3):587–631, September.

Weiwei Sun. 2010. Semantics-driven shallow parsing for chinese semantic role labeling. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 103–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ivan Titov and Alexandre Klementiev. 2012. A Bayesian approach to unsupervised semantic role induction. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, April.

Nianwen Xue. 2008. Labeling chinese predicates with semantic roles. *Computational Linguistics*, 34(2):225–255.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 947–953, Stroudsburg, PA, USA. Association for Computational Linguistics.

Beñat Zapirain, Eneko Agirre, Lluís Màrquez, and Mihai Surdeanu. 2013. Selectional preferences for semantic role classification. In *Computational Linguistics*, pages 631–663.

Tao Zhuang and Chengqing Zong. 2010. Joint inference for bilingual semantic role labeling. In *Proceedings of EMNLP 2010*, pages 304–314, Cambridge, MA, October. Association for Computational Linguistics.