Joint Inference for Bilingual Semantic Role Labeling

Tao Zhuang and **Chengqing Zong**
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
{tzhuang, cqzong}@nlpr.ia.ac.cn

Abstract

We show that jointly performing semantic role labeling (SRL) on bitext can improve SRL results on both sides. In our approach, we use monolingual SRL systems to produce argument candidates for predicates in bitext at first. Then, we simultaneously generate SRL results for two sides of bitext using our joint inference model. Our model prefers the bilingual SRL result that is not only reasonable on each side of bitext, but also has more consistent argument structures between two sides. To evaluate the consistency between two argument structures, we also formulate a log-linear model to compute the probability of aligning two arguments. We have experimented with our model on Chinese-English parallel PropBank data. Using our joint inference model, F1 scores of SRL results on Chinese and English text achieve 79.53% and 77.87% respectively, which are 1.52 and 1.74 points higher than the results of baseline monolingual SRL combination systems respectively.

1 Introduction

In recent years, there has been an increasing interest in SRL on several languages. However, little research has been done on how to effectively perform SRL on bitext, which has important applications including machine translation (Wu and Fung, 2009). A conventional way to perform SRL on bitext is performing SRL on each side of bitext separately, as has been done by Fung et al. (2007) on Chinese-English bitext. However, it is very difficult to obtain good SRL results on both sides of bitext in this way. The reason is that even the state-of-the-art SRL systems do not have very high accuracy on both English text (Màrquez et al., 2008; Pradhan et al., 2008; Punyakanok et al., 2008; Toutanova et al., 2008), and Chinese text (Che et al., 2008; Xue, 2008; Li et al., 2009; Sun et al., 2009).

On the other hand, the semantic equivalence between two sides of bitext means that they should have consistent predicate-argument structures. This bilingual argument structure consistency can guide us to find better SRL results. For example, in Figure 1(a), the argument structure consistency can guide us to choose a correct SRL result on Chinese side. Consistency between two argument structures is reflected by sound argument alignments between them, as shown in Figure 1(b). Previous research has shown that bilingual constraints can be very helpful for parsing (Burkett and Klein, 2008; Huang et al., 2008). In this paper, we show that the bilingual argument structure consistency can be leveraged to substantially improve SRL results on both sides of bitext.

Formally, we present a joint inference model to preform bilingual SRL. Using automatic word alignment on bitext, we first identify a pair of predicates that align with each other. And we use monolingual SRL systems to produce argument candidates for each predicate. Then, our model jointly generate SRL results for both predicates from their argument candidates, using integer linear programming (ILP) technique. An overview of our approach is shown in Figure 2.

Our joint inference model consists of three components: the source side, the target side, and the ar-

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 304–314,
MIT, Massachusetts, USA, 9-11 October 2010.

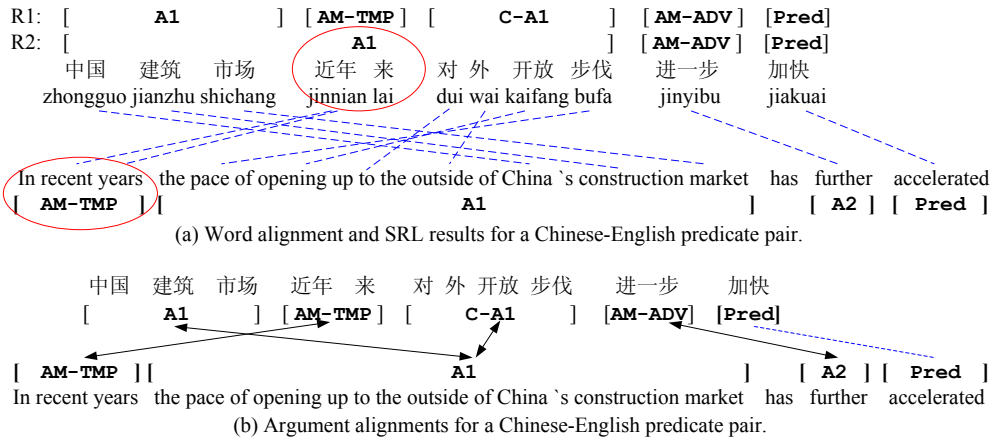

Figure 1: An example from Chinese-English parallel PropBank. In (a), the SRL results are generated by the state-of-the-art monolingual SRL systems. The English SRL result is correct. But it is to more difficult to get correct SRL result on Chinese side, because the AM-TMP argument embeds into a discontinuous A1 argument. The Chinese SRL result in the row marked by 'R1' is correct and consistent with the result on English side. Whereas the result in the row marked by 'R2' is incorrect and inconsistent with the result on English side, with the circles showing their inconsistency. The argument structure consistency can guide us to choose the correct Chinese SRL result.

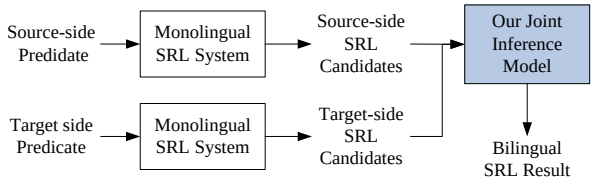

Figure 2: Overview of our approach.

gument alignment between two sides. These three components correspond to three interrelated factors: the quality of the SRL result on source side, the quality of the SRL result on target side, and the argument structure consistency between the SRL results on both sides. To evaluate the consistency between the two argument structures in our joint inference model, we formulate a log-linear model to compute the probability of aligning two arguments. Experiments on Chinese-English parallel PropBank shows that our model significantly outperforms monolingual SRL combination systems on both Chinese and English sides.

The rest of this paper is organized as follows: Section 2 introduces related work. Section 3 describes how we generate SRL candidates on each side of bitext. Section 4 presents our joint inference model. Section 5 presents our experiments. And Section 6 concludes our work.

2 Related Work

Some existing work on monolingual SRL combination is related to our work. Punyakanok et al. (2004; 2008) formulated an ILP model for SRL. Koomen et al. (2005) combined several SRL outputs using ILP method. Màrquez et al. (2005) and Pradhan et al. (2005) proposed combination strategies that are not based on ILP method. Surdeanu et al. (2007) did a complete research on a variety of combination strategies. Zhuang and Zong (2010) proposed a minimum error weighting combination strategy for Chinese SRL combination.

Research on SRL utilizing parallel corpus is also related to our work. Padó and Lapata (2009) did research on cross-lingual annotation projection on English-German parallel corpus. They performed SRL only on the English side, and then mapped the English SRL result to German side. Fung et al. (2007) did pioneering work on studying argument alignment on Chinese-English parallel PropBank. They performed SRL on Chinese and English sides separately. Then, given the SRL result on both sides, they automatically induced the argument alignment between two sides.

The major difference between our work and all existing research is that our model performs SRL inference on two sides of bitext simultaneously. In our

model, we jointly consider three interrelated factors: SRL result on the source side, SRL result on the target side, and the argument alignment between them.

3 Generating Candidates for Inference

3.1 Monolingual SRL System

As shown in Figure 2, we need to use a monolingual SRL system to generate candidates for our joint inference model. We have implemented a monolingual SRL system which utilize full phrase-structure parse trees to perform SRL. In this system, the whole SRL process is comprised of three stages: pruning, argument identification, and argument classification. In the pruning stage, the heuristic pruning method in (Xue, 2008) is employed. In the argument identification stage, a number of argument locations are identified in a sentence. In the argument classification stage, each location identified in the previous stage is assigned a semantic role label. Maximum entropy classifier is employed for both the argument identification and classification tasks. And Zhang Le's MaxEnt toolkit[1] is used for implementation.

We use the monolingual SRL system described above for both Chinese and English SRL tasks. For the Chinese SRL task, the features used in this paper are the same with those used in (Xue, 2008). For the English SRL task, the features used are the same with those used in (Pradhan et al., 2008).

3.2 Output of the Monolingual SRL System

The maximum entropy classifier in our monolingual SRL system can output classification probabilities. We use the classification probability of the argument classification stage as an argument's probability. As illustrated in Figure 3, in an individual system's output, each argument has three attributes: its location in sentence loc, represented by the number of its first word and last word; its semantic role label l; and its probability p.

So each argument outputted by a system is a triple (loc, l, p). For example, the A0 argument in Figure 3 is $((0, 2), \texttt{A0}, 0.94)$. Because these outputs are to be combined, we call such triple a **candidate**.

[1] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Sent:	外商	投资	企业	成为	中国	外贸	重要	增长点
Args:	[A0	]	[Pred]	[		A1		]
loc:	(0, 2)					(4, 7)		
l:	A0					A1		
p:	0.94					0.92		

Figure 3: Three attributes of an output argument: location loc, label l, and probability p.

3.3 Generating and Merging Candidates

To generate candidates for joint inference, we need to have multiple SRL results on each side of bitext. Therefore, for both Chinese and English SRL systems, we use the 3-best parse trees of Berkeley parser (Petrov and Klein, 2007) and 1-best parse trees of Bikel parser (Bikel, 2004) and Stanford parser (Klein and Manning, 2003) as inputs. All the three parsers are multilingual parsers. The second and third best parse trees of Berkeley parser are used for their good quality. Therefore, each monolingual SRL system produces 5 different outputs.

Candidates from different outputs may have the same loc and l but different p. So we merge all candidates with the same loc and l into one by averaging their probabilities. For a merged candidate (loc, l, p), we say that p is **the probability of assigning l to loc.**

4 Joint Inference Model

Our model can be conceptually decomposed to three components: the source side, the target side, and the argument alignment. The objective function of our joint inference model is the weighted sum of three sub-objectives:

$$\max\ O_s + \lambda_1 O_t + \lambda_2 O_a \qquad (1)$$

where O_s and O_t represent the quality of the SRL results on source and target sides, and O_a represents the soundness of the argument alignment between the SRL results on two sides, λ_1, λ_2 are positive weights corresponding to the importance of O_t and O_a respectively.

4.1 Components of Source and Target Sides

4.1.1 Source Side Component

The source side component aims to improve the SRL result on source side. This is equivalent to a

monolingual SRL combination problem.

For convenience, we denote the whole semantic role label set for source language as $\{l_1^s, l_2^s, \ldots, l_{L_s}^s\}$, in which $l_1^s \sim l_6^s$ stand for the key argument labels A0 $\sim$ A5 respectively. Suppose there are N_s different locations, denoted as $loc_1^s, \ldots, loc_{N_s}^s$, among all candidates on the source side. The probability of assigning l_j^s to loc_i^s is p_{ij}^s. An indicator variable x_{ij} is defined as:

$$x_{ij} = [loc_i^s \text{ is assigned label } l_j^s].$$

Then the source side sub-objective O_s in equation (1) is the sum of arguments' probabilities on source side:

$$O_s = \sum_{i=1}^{N_s} \sum_{j=1}^{L_s} (p_{ij}^s - T_s) x_{ij} \quad (2)$$

where T_s is a bias to prevent including too many candidates in solution (Surdeanu et al., 2007).

We consider the following two linguistically motivated constraints:

1. *No duplication*: There is no duplication for key arguments: A0 $\sim$ A5.

2. *No overlapping*: Arguments cannot overlap with each other.

In (Punyakanok et al., 2004), several more constraints are considered. According to (Surdeanu et al., 2007), however, no significant performance improvement can be obtained by considering more constraints than the two above. So we do not consider other constraints.

The inequalities in (3) make sure that each loc_i^s is assigned at most one label.

$$\forall 1 \le i \le N_s : \sum_{j=1}^{L_s} x_{ij} \le 1 \quad (3)$$

The inequalities in (4) satisfy the *No duplication* constraint.

$$\forall 1 \le j \le 6 : \sum_{i=1}^{N_s} x_{ij} \le 1 \quad (4)$$

For any source side location loc_i^s, let C_i denote the index set of the locations that overlap with it. Then the *No overlapping* constraint means that if loc_i^s is assigned a label, i.e., $\sum_{j=1}^{N_s} x_{ij} = 1$, then for any $u \in C_i$, loc_u^s cannot be assigned any label, i.e., $\sum_{j=1}^{N_s} x_{uj} = 0$. A common technique in ILP modeling to form such a constraint is to use a sufficiently large auxiliary constant M. And the constraint is formulated as:

$$\forall 1 \le i \le N_s : \sum_{u \in C_i} \sum_{j=1}^{L_s} x_{uj} \le (1 - \sum_{j=1}^{L_s} x_{ij}) M \quad (5)$$

In this case, M only needs to be larger than the number of candidates to be combined. In this paper, $M = 500$ is large enough.

4.1.2 Target Side Component

In principle, the target side component of our joint inference model is the same with the source side component.

The whole semantic role label set for target language is denoted by $\{l_1^t, l_2^t, \ldots, l_{L_t}^t\}$. There are N_t different locations, denoted as $loc_1^t, \ldots, loc_{N_t}^t$, among all candidates in the target side. And $l_1^t \sim l_6^t$ stand for the key argument labels A0 $\sim$ A5 respectively. The probability of assigning l_j^t to loc_k^t is p_{kj}^t. An indicator variable y_{kj} is defined as:

$$y_{kj} = [loc_k^t \text{ is assigned label } l_j^t].$$

Then the target side sub-objective O_t in equation (1) is:

$$O_t = \sum_{k=1}^{N_t} \sum_{j=1}^{L_t} (p_{kj}^t - T_t) y_{kj} \quad (6)$$

The constraints on target side are as follows:

Each loc_k^t is assigned at most one label:

$$\forall 1 \le k \le N_t : \sum_{j=1}^{L_t} y_{kj} \le 1 \quad (7)$$

The *No duplication* constraint:

$$\forall 1 \le j \le 6 : \sum_{k=1}^{N_t} y_{kj} \le 1 \quad (8)$$

The *No overlapping* constraint:

$$\forall 1 \le k \le N_t : \sum_{v \in C_k} \sum_{j=1}^{L_t} y_{vj} \le (1 - \sum_{j=1}^{L_t} y_{kj}) M \quad (9)$$

In (9), C_k denotes the index set of the locations that overlap with loc_k^t, and the constant M is set to 500 in this paper.

4.2 Argument Alignment

The argument alignment component is the core of our joint inference model. It gives preference to the bilingual SRL results that have more consistent argument structures.

For a source side argument $arg_i^s = (loc_i^s, l^s)$ and a target side argument $arg_k^t = (loc_k^t, l^t)$, let z_{ik} be the following indicator variable:

$$z_{ik} = [arg_i^s \text{ aligns with } arg_k^t].$$

We use p_{ik}^a to represent the probability that arg_i^s and arg_k^t align with each other, i.e., $p_{ik}^a = P(z_{ik} = 1)$. We call p_{ik}^a the **argument alignment probability between** arg_i^s **and** arg_k^t.

4.2.1 Argument Alignment Probability Model

We use a log-linear model to compute the argument alignment probability p_{ik}^a between arg_i^s and arg_k^t. Let (s, t) denote a bilingual sentence pair and wa denote the word alignment on (s, t). Our log-linear model defines a distribution on z_{ik} given the tuple $tup = (arg_i^s, arg_k^t, wa, s, t)$:

$$P(z_{ik}|tup) \propto \exp(w^T \phi(tup))$$

where $\phi(tup)$ is the feature vector. With this model, p_{ik}^a can be computed as $p_{ik}^a = P(z_{ik} = 1|tup)$.

In order to study the argument alignment in corpus and to provide training data for our log-linear model, we have manually aligned the arguments in 60 files (chtb_0121.fid to chtb_0180.fid) of Chinese-English parallel PropBank. On this data set, we get the argument alignment matrix in Table 1.

Ch\En	A0	A1	A2	A3	A4	AM⋆	NUL
A0	492	30	4	0	0	0	46
A1	98	853	43	2	0	0	8
A2	9	57	51	1	0	47	0
A3	1	0	2	6	0	0	0
A4	0	0	2	0	3	0	0
AM⋆	0	2	39	0	0	895	221
NUL	53	14	27	0	0	45	0

Table 1: The argument alignment matrix on manually aligned corpus.

Each entry in Table 1 is the number of times for which one type of Chinese argument aligns with one type of English argument. AM⋆ stands for all adjuncts types like AM-TMP, AM-LOC, etc., and NUL means that the argument on the other side cannot be aligned with any argument on this side. For example, the number 46 in the A0 row and NUL column means that Chinese A0 argument cannot be aligned with any argument on English side for 46 times in our manually aligned corpus.

We use the following features in our model.

Word alignment feature: If there are many word-to-word alignments between arg_i^s and arg_k^t, then it is very probable that arg_i^s and arg_k^t would align with each other. We adopt the method used in (Padó and Lapata, 2009) to measure the word-to-word alignments between arg_i^s and arg_k^t. And the word alignment feature is defined as same as the *word alignment-based word overlap* in (Padó and Lapata, 2009). Note that this is a real-valued feature.

Head word alignment feature: The head word of an argument is usually more representative than other words. So we use whether the head words of arg_i^s and arg_k^t align with each other as a binary feature. The use of this feature is inspired by the work in (Burkett and Klein, 2008).

Semantic role labels of two arguments: From Table 1, we can see that semantic role labels of two arguments are a good indicator of whether they should align with each other. For example, a Chinese A0 argument aligns with an English A0 argument most of the times, and never aligns with an English AM⋆ argument in Table 1. Therefore, the semantic role labels of arg_i^s and arg_k^t are used as a feature.

Predicate verb pair: Different predicate pairs have different argument alignment patterns. Let's take the Chinese predicate 增长/*zengzhang* and the English predicate *grow* as an example. The argument alignment matrix for all instances of the Chinese-English predicate pair (*zengzhang*, *grow*) in our manually aligned corpus is shown in Table 2.

CH \EN	A0	A1	A2	AM⋆	NUL
A0	0	16	0	0	0
A1	0	0	12	0	0
AM⋆	0	0	4	7	10
NUL	0	0	0	2	0

Table 2: The argument alignment matrix for the predicate pair (*zengzhang*, *grow*).

From Table 2 we can see that all A0 arguments of *zengzhang* align with A1 arguments of *grow*. This

is very different from the results in Table 1, where a Chinese A0 argument tends to align with an English A0 argument. This phenomenon shows that a predicate pair can determine which types of arguments should align with each other. Therefore, we use the predicate pair as a feature.

4.2.2 Argument Alignment Component

The argument alignment sub-objective O_a in equation (1) is the sum of argument alignment probabilities:

$$O_a = \sum_{i=1}^{N_s}\sum_{k=1}^{N_t}(p_{ik}^a - T_a)z_{ik} \quad (10)$$

where T_a is a bias to prevent including too many alignments in final solution, and p_{ik}^a is computed using the log-linear model described in subsection 4.2.1.

O_a reflects the consistency between argument structures on two sides of bitext. Larger O_a means better argument alignment between two sides, thus indicates more consistency between argument structures on two sides.

The following constraints are considered:

1. *Conformity with bilingual SRL result*. For all candidates on both source and target sides, only those that are chosen to be arguments on each side can be aligned.

2. *One-to-many alignment limit*. Each argument can not be aligned with more than 3 arguments.

3. *Complete argument alignment*. Each argument on source side must be aligned with at least one argument on target side, and vice versa.

The *Conformity with bilingual SRL result* constraint is necessary to validly integrate the bilingual SRL result with the argument alignment. This constraint means that if arg_i^s and arg_k^t align with each other, i.e., $z_{ik} = 1$, then loc_i^s must be assigned a label on source side, i.e., $\sum_{j=1}^{L_s} x_{ij} = 1$, and loc_k^t must be assigned a label on target side, i.e., $\sum_{j=1}^{L_t} y_{kj} = 1$. So this constraint can be represented as:

$$\forall 1 \le i \le N_s, 1 \le k \le N_t : \sum_{j=1}^{L_s} x_{ij} \ge z_{ik} \quad (11)$$

$$\forall 1 \le k \le N_t, 1 \le i \le N_s : \sum_{j=1}^{L_t} y_{kj} \ge z_{ik} \quad (12)$$

The *One-to-many alignment limit* constraint comes from our observation on manually aligned corpus. We have found that no argument aligns with more than 3 arguments in our manually aligned corpus. This constraint can be represented as:

$$\forall 1 \le i \le N_s : \sum_{k=1}^{N_t} z_{ik} \le 3 \quad (13)$$

$$\forall 1 \le k \le N_t : \sum_{i=1}^{N_s} z_{ik} \le 3 \quad (14)$$

The *Complete argument alignment* constraint comes from the semantic equivalence between two sides of bitext. For each source side location loc_i^s, if it is assigned a label, i.e., $\sum_{j=1}^{L_s} x_{ij} = 1$, then it must be aligned with some arguments on target side, i.e., $\sum_{k=1}^{N_t} z_{ik} \ge 1$. This can be represented as:

$$\forall 1 \le i \le N_s : \sum_{k=1}^{N_t} z_{ik} \ge \sum_{j=1}^{L_s} x_{ij} \quad (15)$$

Similarly, each target side argument must be aligned to at least one source side argument. This can be represented as:

$$\forall 1 \le k \le N_t : \sum_{i=1}^{N_s} z_{ik} \ge \sum_{j=1}^{L_t} y_{kj} \quad (16)$$

4.3 Complete Argument Alignment as a Soft Constraint

Although the hard *Complement argument alignment* constraint is ideally reasonable, in real situations this constraint does not always hold. The manual argument alignment result shown in Table 1 indicates that in some cases an argument cannot be aligned with any argument on the other side (see the NUL row and column in Table 1). Therefore, it would be reasonable to change the hard *Complement argument alignment* constraint to a soft one. To do so, we need to remove the hard *Complement argument alignment* constraint and add penalty for violation of this constraint.

If an argument does not align with any argument on the other side, we say it aligns with NUL. And we define the following indicator variables:
$z_{i,NUL} = [arg_i^s \text{ aligns with NUL}], 1 \le i \le N_s$.

$z_{NUL,k} = [arg_k^t$ aligns with NUL$], 1 \leq k \leq N_t$. Then $\sum_{i=1}^{N_s} z_{i,NUL}$ is the number of source side arguments that align with NUL. And $\sum_{k=1}^{N_t} z_{NUL,k}$ is the number of target side arguments that align with NUL. For each argument that aligns with NUL, we add a penalty λ_3 to the argument alignment sub-objective O_a. Therefore, the sub-objective O_a in equation (10) is changed to:

$$O_a = \sum_{i=1}^{N_s}\sum_{k=1}^{N_t}(p_{ik}^a - T_a)z_{ik} - \lambda_3(\sum_{i=1}^{N_s} z_{i,NUL} + \sum_{k=1}^{N_t} z_{NUL,k}) \quad (17)$$

From the definition of $z_{i,NUL}$, it is obvious that, for any $1 \leq i \leq N_s$, $z_{i,NUL}$ and $z_{ik}(1 \leq k \leq N_t)$ have the following relationship: If $\sum_{k=1}^{N_t} z_{ik} \geq 1$, i.e., arg_i^s aligns with some arguments on target side, then $z_{i,NUL} = 0$; Otherwise, $z_{i,NUL} = 1$. These relationships can be captured by the following constraints:

$$\forall 1 \leq i \leq N_s, 1 \leq k \leq N_t : z_{i,NUL} \leq 1 - z_{ik} \quad (18)$$

$$\forall 1 \leq i \leq N_s : \sum_{k=1}^{N_t} z_{ik} + z_{i,NUL} \geq 1 \quad (19)$$

Similarly, for any $1 \leq k \leq N_t$, $z_{NUL,k}$ and $z_{ik}(1 \leq i \leq N_s)$ observe the following constraints:

$$\forall 1 \leq k \leq N_t, 1 \leq i \leq N_s : z_{NUL,k} \leq 1 - z_{ik} \quad (20)$$

$$\forall 1 \leq k \leq N_t : \sum_{i=1}^{N_s} z_{ik} + z_{NUL,k} \geq 1 \quad (21)$$

4.4 Models Summary

So far, we have presented two versions of our joint inference model. The first version treats *Complement argument alignment* as a hard constraint. We will refer to this version as *Joint1*. The objective function of *Joint1* is defined by equations (1, 2, 6, 10). And the constraints of *Joint1* are defined by equations (3-5, 7-9, 11-16).

The sencond version treats *Complement argument alignment* as a soft constraint. We will refer to this version as *Joint2*. The objective function of *Joint2* is defined by equations (1, 2, 6, 17). And the constraints of *Joint2* are defined by equations (3-5, 7-9, 11-14, 18-21).

Our baseline models are monolingual SRL combination models. We will refer to the source side combination model as *SrcCmb*. The objective of *SrcCmb* is to maximize O_s, which is defined in equation (2). And the constraints of *SrcCmb* are defined by equations (3-5). Similarly, we will refer to the target side combination model as *TrgCmb*. The objective of *TrgCmb* is to maximize O_t defined in equation (6). And the constraints of *TrgCmb* are defined by equations (7-9). In this paper, we employ lpsolve[2] to solve all ILP models.

5 Experiments

5.1 Experimental Setup

In our experiments, we use the Xinhua News portion of Chinese and English data in LDC OntoNotes Release 3.0. This data is a Chinese-English parallel proposition bank described in (Palmer et al., 2005). It contains parallel proposition annotations for 325 files (chtb_0001.fid to chtb_0325.fid) from Chinese-English parallel Treebank. The English part of this data contains proposition annotations only for verbal predicates. Therefore, we only consider verbal predicates in this paper.

We employ the GIZA++ toolkit (Och and Ney, 2003) to perform automatic word alignment. Besides the parallel PropBank data, we use additional 4,500K Chinese-English sentence pairs[3] to induce word alignments for both directions, with the default GIZA++ settings. The alignments are symmetrized using the intersection heuristic (Och and Ney, 2003), which is known to produce high-precision alignments.

We use 80 files (chtb_0001.fid to chtb_0080.fid) as test data, and 40 files (chtb_0081.fid to chtb_0120.fid) as development data. Although our joint inference model needs no training, we still need to train a log-linear argument alignment probability model, which is used in the joint inference model. As specified in subsection 4.2.1, the train-

[2] http://lpsolve.sourceforge.net/

[3] These data includes the following LDC corpus: LDC2002E18, LDC2003E07, LDC2003E14, LDC2005T06, LDC2004T07, LDC2000T50.

ing set for the argument alignment probability model consists of 60 files (chtb_0121.fid to chtb_0180.fid) with manual argument alignment. Unfortunately, the quality of automatic word alignment on one-to-many Chinese-English sentence pairs is usually very poor. So we only include one-to-one Chinese-English sentence pairs in all data. And not all predicates in a sentence pair can be included. Only bilingual predicate pairs are included. A bilingual predicate pair is defined to be a pair of predicates in bitext which align with each other in automatic word alignment. Table 3 shows how many sentences and predicates are included in each data set.

	Test	Dev	Train
Articles	1-80	81-120	121-180
Chinese Sentences	1067	578	778
English Sentences	1182	620	828
Bilingual pairs	821	448	614
Chinese Predicates	3792	2042	2572
English Predicates	2864	1647	1860
Bilingual pairs	1476	790	982

Table 3: Sentence and predicate counts.

Our monolingual SRL systems are trained separately. Our Chinese SRL system is trained on 640 files (chtb_0121.fid to chtb_0931.fid) in Chinese Propbank 1.0. Because Xinhua News is a quite different domain from WSJ, the training set for our English SRL system includes not only Sections 02∼21 of WSJ data in English Propbank, but also 205 files (chtb_0121.fid to chtb_0325.fid) in the English part of parallel PropBank. For Chinese, the syntactic parsers are trained on 640 files (chtb_0121.fid to chtb_0931.fid) plus the broadcast news portion of Chinese Treebank 6.0. For English, the syntactic parsers are trained on the following data: Sections 02∼21 of WSJ data in English Treebank, 205 files (chtb_0121.fid to chtb_0325.fid) of Xinhua News data in OntoNotes 3.0, and the Sinorama data in OntoNotes 3.0. We treat discontinuous and coreferential arguments in accordance to the CoNLL-2005 shared task (Carreras and Màrquez, 2005). The first part of a discontinuous argument is labeled as it is, and the second part is labeled with a prefix "C-". All coreferential arguments are labeled with a prefix "R-".

5.2 Tuning Parameters in Models

The models *Joint1*, *Joint2*, *SrcCmb*, and *TrgCmb* have different parameters. For each model, we have automatically tuned its parameters on development set using Powell's Mothod (Brent, 1973). Powell's Method is a heuristic optimization algorithm that does not require the objective function to have an explicit analytical formula. For a monolingual model like *SrcCmb* or *TrgCmb*, our objective is to maximize the F_1 score of the model's result on development set. But a joint model, like *Joint1* or *Joint2*, generates SRL results on both sides of bitext. So our objective is to maximize the sum of the two F_1 scores of the model's results for both Chinese and English on development set. For all models, we regard the parameters to be tuned as variables. Then we optimize our objective using Powell's Method. The solution of this optimization is the values of parameters. To avoid finding poor local optimum, we perform the optimization 30 times with different initial parameter values, and choose the best solution found. The final parameter values are listed in Table 4.

Model	T_s	T_t	T_a	λ_1	λ_2	λ_3
SrcCmb	0.21	-	-	-	-	-
TrgCmb	-	0.32	-	-	-	-
Joint1	0.17	0.22	0.36	0.96	1.04	-
Joint2	0.15	0.26	0.42	1.02	1.21	0.15

Table 4: Parameter values in models.

5.3 Individual SRL Outputs' Performance

As specified in subsection 3.3, the monoligual SRL system uses different parse trees to generate multiple SRL outputs. The performance of these outputs on test set is shown in Table 5. In Table 5, O1∼O3 are the outputs using 3-best parse trees of Berkeley parser respectively, O4 and O5 are the outputs using the best parse trees of Stanford parser and Bikel parser respectively.

As specified in subsection 5.1, only a small part of English SRL training data is in the same domain with test data. Therefore, the English SRL result in Table 5 is not very impressive. But the Chinese SRL result is pretty good.

Side	Outputs	P(%)	R(%)	F_1
Chinese	O1	**79.84**	**71.95**	**75.69**
	O2	78.53	70.32	74.20
	O3	78.41	69.99	73.96
	O4	73.21	67.13	70.04
	O5	75.32	63.78	69.07
English	O1	**77.13**	**70.42**	**73.62**
	O2	75.88	69.06	72.31
	O3	75.74	68.65	72.02
	O4	71.57	66.11	68.73
	O5	73.12	68.04	70.49

Table 5: The results of individual monolingual SRL outputs on test set.

5.4 Effects of Different Constraints

The *One-to-many limit* and *Complete argument alignment* constraints in subsection 4.2.2 comes from our empirical knowledge. To investigate the effect of these two constraits, we remove them from our joint inference models one by one, and observe the performance variations on test set. The results are shown in Table 6. In Table 6, 'c2' refers to the *One-to-many limit* constraint, 'c3' refers to the *Complete argument alignment* constraint, and '-' means removing. For example, '*Joint1* - c2' means removing the constraint 'c2' from the model *Joint1*. Recall that the only difference between *Joint1* and *Joint2* is that 'c3' is a hard constraint in *Joint1*, but a soft constraint in *Joint2*. Therefore, '*Joint2* - c3' and '*Joint2* - c2 - c3' do not appear in Table 6, because they are the same with '*Joint1* - c3' and '*Joint1* - c2 - c3' respectively.

Model	Side	P(%)	R(%)	F_1
Joint1	Chinese	82.95	75.21	78.89
Joint1 - c2		81.46	75.97	78.62
Joint1 - c3		82.36	74.68	78.33
Joint1 - c2 - c3		82.04	74.67	78.18
Joint2		**83.35**	**76.04**	**79.53**
Joint2 - c2		82.41	76.03	79.09
Joint1	English	79.38	75.16	77.21
Joint1 - c2		78.51	75.22	76.83
Joint1 - c3		78.66	74.55	76.55
Joint1 - c2 - c3		78.37	74.37	76.32
Joint2		**79.64**	**76.18**	**77.87**
Joint2 - c2		78.41	75.89	77.13

Table 6: Results of different joint models on test set.

From Table 6, we can see that the constraints 'c2' and 'c3' both have positive effect in our joint inference model, because removing any one of them causes performance degradation. And removing 'c3' from *Joint1* causes more performance degradation than removing 'c2'. This means that 'c3' plays a more important role than 'c2' in our joint inference model. Indeed, by treating 'c3' as a soft constraint, the model *Joint2* has the best performance on both sides of bitext.

5.5 Final Results

We use *Joint2* as our final joint inference model. And as specified in subsection 4.4, our baselines are monolingual SRL combination models: *SrcCmb* for Chinese, and *TrgCmb* for English. Note that *SrcCmb* and *TrgCmb* are basically the same as the state-of-the-art combination model in (Surdeanu et al., 2007) with *No overlapping* and *No duplication* constraints. The final results on test set are shown in Table 7.

Side	Model	P(%)	R(%)	F_1
Chinese	*SrcCmb*	82.58	73.92	78.01
	Joint2	**83.35**	**76.04**	**79.53**
English	*TrgCmb*	79.02	73.44	76.13
	Joint2	**79.64**	**76.18**	**77.87**

Table 7: Comparison between monolingual combination model and our joint inference model on test set.

From Table 5 and Table 7, we can see that *SrcCmb* and *TrgCmb* improve F_1 scores over the best individual SRL outputs by 2.32 points and 2.51 points on Chinese and English seperately. Thus they form strong baselines for our joint inference model. Even so, our joint inference model still improves F_1 score over *SrcCmb* by 1.52 points, and over *TrgCmb* by 1.74 points.

From Table 7, we can see that, despite only part of training data for English SRL system is in-domain, our joint inference model still produces good English SRL result. And the F_1 score of Chinese SRL result reaches 79.53%, which represents the state-of-the-art Chinese SRL performance to date.

6 Conclusions

In this paper, we propose a joint inference model to perform bilingual SRL. Our joint inference model incorporates not only linguistic constraints on

source and target sides of bitext, but also the bilingual argument structure consistency requirement on bitext. Experiments on Chinese-English parallel PropBank show that our joint inference model is very effective for bilingual SRL. Compared to state-of-the-art monolingual SRL combination baselines, our joint inference model substantially improves SRL results on both sides of bitext. In fact, the solution of our joint inference model contains not only the SRL results on bitext, but also the optimal argument alignment between two sides of bitext. This makes our model especially suitable for application in machine translation, which needs to obtain the argument alignment.

Acknowledgments

The research work has been partially funded by the Natural Science Foundation of China under Grant No. 60975053 and 60736014, the National Key Technology R&D Program under Grant No. 2006BAH03B02. We would like to thank Jiajun Zhang for helpful discussions and the anonymous reviewers for their valuable comments.

References

Daniel Bikel. 2004. Intricacies of Collins Parsing Model. *Computational Linguistics*, 30(4):480-511.

Richard P. Brent. 1973. *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

David Burkett, and Dan Klein. 2008. Two Languages are Better than One (for Syntactic Parsing). In *Proceedings of EMNLP-2008*, pages 877-886.

Xavier Carreras, and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: semantic role labeling. In *Proceedings of CoNLL-2005*, pages 152-164.

Wanxiang Che, Min Zhang, Ai Ti Aw, Chew Lim Tan, Ting Liu, and Sheng Li. 2008. Using a Hybrid Convolution Tree Kernel for Semantic Role Labeling. *ACM Transactions on Asian Language Information Processing*, 2008, 7(4): 1-23.

Pascale Fung, Zhaojun Wu, Yongsheng Yang and Dekai Wu. 2007. Learning Bilingual Semantic Frames: Shallow Semantic Parsing vs. Semantic Role Projection. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation*, pages 75-84.

Liang Huang, Wenbin Jiang, Qun Liu. 2009. Bilingually-Constrained (Monolingual) Shift-Reduce Parsing. In *Proceedings of EMNLP-2009*, pages 1222-1231.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL-2003*, pages 423-430.

Peter Koomen, Vasin Punyakanok, Dan Roth, and Wentau Yih. 2005. Generalized Inference with Multiple Semantic Role Labeling Systems. In *Proceedings of CoNLL-2005 shared task*, pages 181-184.

Junhui Li, Guodong Zhou, Hai Zhao, Qiaoming Zhu, and Peide Qian. 2009. Improving Nominal SRL in Chinese Language with Verbal SRL Information and Automatic Predicate Recognition. In *Proceedings of EMNLP-2009*, pages 1280-1288.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, Suzanne Stevenson. 2008. Semantic Role Labeling: An Introduction to the Special Issue. *Computational Linguistics*, 34(2):145-159.

Lluís Màrquez, Mihai Surdeanu, Pere Comas, and Jordi Turmo. 2005. A Robust Combination Strategy for Semantic Role Labeling. In *Proceedings of EMNLP-2005*, pages 644-651.

Frans J. Och, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19-51.

Sebastian Padó, and Mirella Lapata. 2009. Cross-lingual Annotation Projection of Semantic Roles. *Journal of Artificial Intelligence Research (JAIR)*, 36:307-340.

Martha Palmer, Nianwen Xue, Olga Babko-Malaya, Jinying Chen, Benjamin Snyder. 2005. A Parallel Proposition Bank II for Chinese and English. In *Frontiers in Corpus Annotation, Workshop in conjunction with ACL-05*, pages 61-67.

Slav Petrov and Dan Klein. 2007. Improved Inference for Unlexicalized parsing. In *Proceedings of ACL-2007*, pages 46-54.

Sameer S. Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Daniel Jurafsky. 2005. Semantic Role Labeling Using Different Syntactic Views. In *Proceedings of ACL-2005*, pages 581-588.

Sameer S. Pradhan, Wayne Ward, James H. Martin. 2008. Towards Robust Semantic Role Labeling. *Computational Linguistics*, 34(2):289-310.

Vasin Punyakanok, Dan Roth, Wen-tauYih. 2008. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics*, 34(2):257-287.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. 2004. Semantic Role Labeling via Integer Linear Programming Inference. In *Proceedings of COLING-2004*, pages 1346-1352.

Weiwei Sun, Zhifang Sui, Meng Wang, and Xin Wang. 2009. Chinese Semantic Role Labeling with Shallow

Parsing. In *Proceedings of EMNLP-2009*, pages 1475-1483.

Mihai Surdeanu, Lluís Màrquez, Xavier Carreras, and Pere R. Comas. 2007. Combination Strategies for Semantic Role Labeling. *Journal of Artificial Intelligence Research (JAIR)*, 29:105-151.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. 2008. A Global Joint Model for Semantic Role Labeling. *Computational Linguistics*, 34(2): 145-159.

Dekai Wu, and Pascale Fung. 2009. Semantic Roles for SMT: A Hybrid Two-Pass Model. In *Proceedings of NAACL-2009*, pages 13-16.

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

Tao Zhuang, and Chengqing Zong. 2010. A Minimum Error Weighting Combination Strategy for Chinese Semantic Role Labeling. In *Proceedings of COLING-2010*, pages 1362-1370.