Improving Chinese-English PropBank Alignment

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 2 improvements to Chinese-English PropBank predicate-argument structure alignment. Taking advantage of the recently expanded PropBank English nominal and adjective predicate annotation (Bonial et al., 2014), we performed predicate-argument alignments between both verb and nominal/adjective predicates in Chinese and English. Using our alignment system, this increased the number of aligned predicate-argument structures by 24.5% on the parallel Xinhua News corpus. We also improved the PropBank alignment system using expectation-maximization (EM) techniques. By collecting Chinese-English predicate-to-predicate and argument type-to-argument type alignment probabilities and iteratively improving the alignment output using these probabilities on a large unannotated parallel corpora, we improved the predicate alignment performance by 1 F point when using all automatic SRL and word alignment inputs.

1 Introduction

With the growing interest in building semantically-driven machine translation (MT) systems/evaluation metrics (Carpuat and Wu, 2007; Wu and Fung, 2009b; Wu and Fung, 2009a; Lo and Wu, 2011; Lo et al., 2013; Ma, 2014), the need for a comprehensive and high performing semantic alignment system has become more pressing. While there are finer grained representations such as FrameNet (Baker et al., 1998) and Abstract Meaning Representation (AMR) (Banarescu et al., 2013), PropBank (Palmer et al., 2005) semantic representation has been popular in the MT community partly because of the availability of large quantity of annotated data in multiple languages, enabling the development of accurate automatic semantic role labeling systems.

arg type	Arg0	Arg1	Arg2	Arg3	Arg4	V
Arg0	***1610***	79	25	-	-	9
Arg1	432	***2665***	128	**11**	-	142
Arg2	43	*310*	**140**	8	3	67
Arg3	2	14	*21*	7	-	4
Arg4	1	*37*	9	3	**6**	4
V	25	28	22	1	-	**3278**

Table 1: Chinese argument type (column) to English argument type (row) alignment counts using gold SRL and word alignment annotated Xinhua News data

While the argument types defined in PropBank were intended to be self-contained and independent of the predicate or language, as Fung et al. (2007), Choi et al. (2009), and our previous work (Wu and Palmer, 2011) have demonstrated, assuming alignment between arguments of the same type is insufficient. Table 1 shows the alignment distribution of the core argument types between Chinese and English. While ARG0 and ARG1 alignments are relatively deterministic, alignment involving ARG2–5 and adjunct argument types (not shown) are much more varied. Part of this alignment variety is caused by differences in argument annotation guidelines between English and Chinese, but another part is caused by verb predicates being nominalized in the translation. Our previous work tried to address the first issue by using aligned words in the argument span (instead of the argument type) to align argu-

Proceedings of SSST-9, Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 74–82,
Denver, Colorado, June 4, 2015.

ments between English and Chinese. But since we had only considered alignments between verb predicates and their arguments, around 27% of the time, verb predicates are aligned somewhat awkwardly. Another issue we had encountered is, since we solely relied on word alignment input, the approach is not very reliable for aligning short arguments (since a single word alignment error can become critical).

In this work, we attempt to address both of these issues. With the recently expanded PropBank English nominal and adjective predicate annotation (Bonial et al., 2014), we are now able to perform predicate-argument alignments between both verb and nominal/adjective predicates in Chinese and English. With our alignment system, this increased the number of aligned predicate-argument structures by 24.5% on the parallel Xinhua News corpus and allowed more semantically similar predicates to be aligned, regardless of the syntactic form of the predicates. We also propose an extension to our predicate-argument alignment system by factoring in predicate-to-predicate and argument type-to-argument type alignment probabilities when making alignment decisions. Combined with expectation-maximization (EM) techniques that iteratively refines these probabilities, we achieved an 1 F1 point predicate alignment performance improvement using all automatic (SRL and word alignment) inputs. More over, even though the alignment probabilities were generated from automatic system inputs, in some instances, we were able to improve alignment performances using gold SRL inputs.

2 Related Work

Resnik (2004) was one of the earlier works proposing semantic similarity (with a looser definition of semantically similar/equivalent phrases) using triangulation between parallel corpora. This was extended later by Madnani et al. (2008a; 2008b)). Mareček (2009) proposed aligning tectogrammatical trees, where only content (autosemantic) words are nodes, in a parallel English/Czech corpus to improve overall word alignment and thereby improve machine translation. Padó and Lapata (2005; 2006) used word alignment and syntax based argument similarity to project English FrameNet semantic roles to German.

Fung et al. (2007) demonstrated that there is poor semantic parallelism between Chinese-English bilingual sentences. Their technique for improving Chinese-English predicate-argument mapping ($ARG_{Chinese,i} \mapsto ARG_{English,j}$) consists of matching predicates with a bilingual lexicon, computing cosine-similarity (based on lexical translation) of (only) core arguments and tuning on an unannotated parallel corpus. Choi et al. (2009) showed how to enhance Chinese-English verb alignments by exploring predicate-argument structure alignment using parallel PropBanks. The system, using GIZA++ word alignment, deduced alternate verb alignments that showed improvement over pure GIZA++ alignment.

Wu and Fung (2009b) was one of the first to use parallel semantic roles to improve MT system output. Given the outputs from Moses (Koehn et al., 2007), a machine translation decoder, they reordered the translations based on the best predicate-argument alignment. The resulting system showed a 0.5 point BLEU score improvement even though the BLEU metric often discounts improvement in semantic consistency of MT output. To address this issue, Lo and Wu (2011) proposed MEANT, a predicate-argument structure alignment based machine translation evaluation system that better correlates with human MT judgment. Lo et al. (2013) later showed that tuning an MT system against this metric produced more robust translations. Similar ideas on semantically coherent MT have been explored by Ma (2014), where the system attempts to fuse multiple MT translations using predicate-argument alignment metrics, though the results did not show improvement with the BLEU metric.

More recently, Banarescu et al. (2013) have proposed Abstract Meaning Representation (AMR) as an alternative/intermediary representation for MT that may improve the semantic coherency of the output. While the project have only recently gained more traction, an AMR-based MT would likely require aligning AMR concepts between the 2 translation languages. Since AMR is based to a large degree on PropBank SRL, improving SRL alignment should transfer accordingly to improvements in AMR alignments as well.

3 Aligning PropBank Predicate-Arguments

Given a parallel sentence pair, we attempt to find the corresponding PropBank predicate-argument alignments between the sentences as illustrated by figure 1.

3.1 Baseline approach

We first describe our baseline predicate-argument alignment approach (Wu and Palmer, 2011): argument alignments are based on the proportion of aligned words between them, predicate-argument structure alignments are based on the alignment quality of their arguments. We assume there can be a many-to-many argument alignment but only a one-to-one predicate-argument structure alignment between the 2 languages.

Formally, we denote $a_{i,c}$ and $a_{j,e}$ as arguments in Chinese and English respectively, $A_{I,c}$ and $A_{J,e}$ as a set of mapped Chinese and English arguments respectively, $W_{i,c}$ as the words in argument $a_{i,c}$, and $map_e(a_{i,c}) = W_{i,e}$ as the word alignment function that takes the source argument and produces a set of words in the target language sentence. We define precision as the fraction of aligned target words in the mapped argument set:

$$P_{I,c} = \frac{|(\cup_{i \in I} map_e(a_{i,c})) \cap (\cup_{j \in J} W_{j,e})|}{|\cup_{i \in I} map_e(a_{i,c})|} \quad (1)$$

and recall as the fraction of source words in the mapped argument set:

$$R_{I,c} = \frac{\sum_{i \in I} |W_{i,c}|}{\sum_{\forall i} |W_{i,c}|} \quad (2)$$

We then choose the $A_{I,c}$ that optimizes the F1-score of P_c and R_c:

$$A_{I,c} = \arg\max_{I} \frac{2 \cdot P_{I,c} \cdot R_{I,c}}{P_{I,c} + R_{I,c}} = F_{I,c} \quad (3)$$

Finally, to constrain both the source and target argument sets, we optimize:

$$A_{I,c}, A_{J,c} = \arg\max_{I,J} \frac{2 \cdot F_{I,c} \cdot F_{J,e}}{F_{I,c} + F_{J,e}} = F_{IJ} \quad (4)$$

To measure similarity between a single pair of source, target arguments, we define:

$$P_{ij} = \frac{|map_e(a_{i,c}) \cup W_{j,e}|}{|map_e(a_{i,c})|}$$
$$R_{ij} = \frac{|map_c(a_{j,e}) \cup W_{i,c}|}{|map_c(a_{j,e})|} \quad (5)$$

While our work has demonstrated that this approach can produce better predicate alignments than word alignment alone, it can also become confused when there are multiple predicates in a sentence that have shared words in their argument spans, especially when faced with word alignment errors. Figure 2 shows one such example: because the automatic word aligner erroneously aligned both 自筹/*self-provide* and 建设/*construct* to *build* (shown with dotted lines), as well as missed the correct word alignments of 自筹 to *Using its own*, 自筹 is instead aligned to *build*, since they share more aligned words amongst the arguments. However, since the Chinese predicate 建设/*construct* often aligns to *build*, and `ARG1` in Chinese frequently maps to `ARG1` in English but rarely maps to `ARGM-MNR`, an alignment framework that considers these likelihood can potentially correct these types of misalignment.

3.2 Building a alignment probability model

To enhance our baseline approach, we first collect alignment probabilities between a Chinese predicate and its argument types and a English predicate and its argument types. Specifically, we are interested in the following:

$p(pred_{j,e}|pred_{i,c})$: given a Chinese predicate in the mapping, the probability of an English predicate

$p(a_{l,e}|a_{k,c}, pred_{i,c}, pred_{j,e})$: given an aligned Chinese & English predicate pair and the Chinese argument type, the probability of an English argument type

In addition to producing a better alignment output, these 2 probabilities (along with probabilities in the English-to-Chinese alignment direction) may also be used to compute the semantic similarity of a pair of parallel sentences.

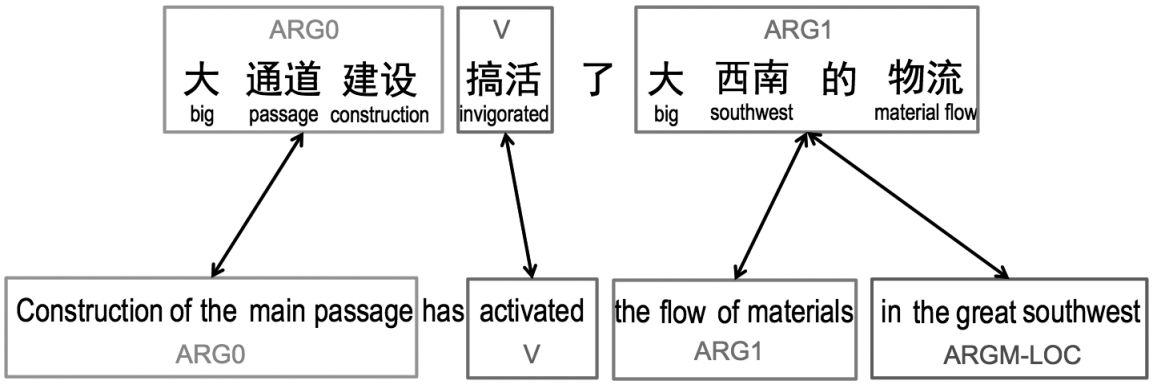

Figure 1: Chinese predicate-arguments mapping example

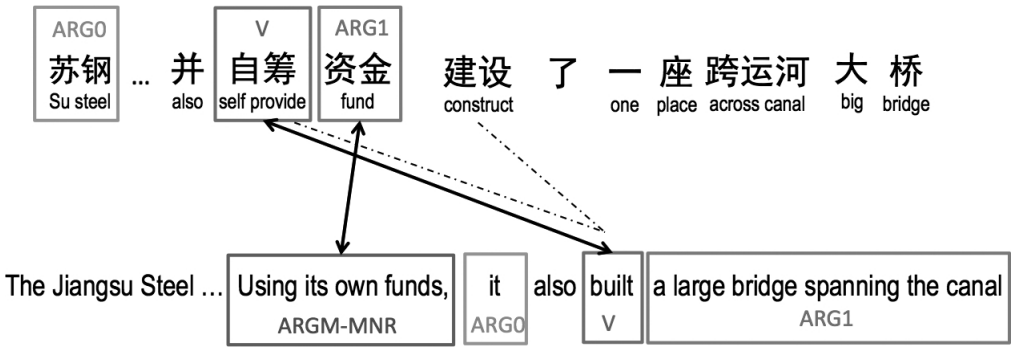

Figure 2: Bad predicate-argument alignment (solid lines) caused by word alignment (dashed lines) error

3.2.1 Predicate-to-predicate mapping probability

There are over 20,000 Chinese predicates and over 10,000 English predicates (in OntoNotes 5.0 PropBank frame files). Even on a large corpora, $freq_{map}(pred_{i,c}, pred_{j,e})$ will be low or zero for many predicate pairs when producing a probability estimate. We chose the Simple Good-Turing smoothing method (Gale, 1995) to smooth the seen mapping frequency counts and estimate the total unseen mapping probability $\sum_{j \in freq_{map}(pred_{i,c}, pred_{j,e})=0} p(pred_{j,e}|pred_{i,c})$.

3.2.2 Argument-to-argument mapping probability

Since $freq_{map}(pred_{j,e}|pred_{i,c})$ is sparse, $freq_{map}(a_{l,e}|pred_{i,c}, pred_{j,e}, a_{k,c})$ will also be sparse. We address this using absolute discounting (Chen and Goodman, 1996) to smooth

$$p(a_{l,e}|a_{k,c}, pred_{i,c}, pred_{j,e}) = \frac{\max(freq(a_{l,e}|a_{k,c}, pred_{i,e}, pred_{j,e}) - d, 0)}{\sum_l freq(a_{l,e}|a_{k,c}, pred_{i,c}, pred_{j,e})} + (1 - \lambda) \cdot p_{backoff}(a_{l,e})$$

with a few different back-off probability distributions:

(a) $p(a_{l,e}|a_{k,c}, pred_{i,c})$: given the Chinese predicate and argument type, the probability of an English argument type

(b) $p(a_{l,e}|a_{k,c}, pred_{j,e})$: given the English predicate and Chinese argument type, the probability of an English argument type

(c) $p(a_{l,e}|a_{k,c})$: given the Chinese argument type, the probability of an English argument type

(a) and (b) can be further smoothed using (c), while (c) can be computed directly from the frequency

count over a large corpus since there are less than 30 argument types for either Chinese or English. To choose between (a) and (b) as the back-off probability distribution, we compute the $cosine$ similarity between (a), (c) and (b), (c) and choose the smaller of the 2 (i.e., choose the more specific distribution that's less similar (more informative) to the base distribution).

3.3 Probabilistic alignment

With the probability model described previously, we attempted to improve predicate-argument alignment by integrating the model with the alignment algorithm. Because the model is computed using automatic system output, we wanted to ensure the alignment algorithm does not overly rely on it. Therefore we modify equation 5 to:

$$\begin{aligned} P'_{kl} =& (1 - \beta + \beta \cdot w(a_{l,e}|a_{k,c}, pred_{i,c}, pred_{j,e}))P_{kl} \\ R'_{kl} =& (1 - \beta + \beta \cdot w(a_{k,c}|a_{l,e}, pred_{i,c}, pred_{j,e}))R_{kl} \end{aligned} \tag{6}$$

where $0 \leq \beta \leq 1$ and

$$w(a_k) = \frac{p(a_k)}{\sum_k p(a_k) \cdot p(a_k)} \tag{7}$$

so that the expected value of $w(a_k)$, $E(w(a_k)) = 1$. If $P'_{kl} > 1$ or $R'_{kl} > 1$, we change $P'_{kl} = 1$, $R'_{kl} = 1$. We also update equation 3 to take into account predicate-to-predicate mapping likelihood:

$$\begin{aligned} F'_{i,c} &= (1 - \alpha + \alpha \cdot w(pred_{j,e}|pred_{i,c}))F_{i,c} \\ F'_{j,e} &= (1 - \alpha + \alpha \cdot w(pred_{i,c}|pred_{j,e}))F_{j,e} \end{aligned} \tag{8}$$

We choose α and β (through grid-search) to maximize the sum of the alignment score of all the predicate-argument pairs in the corpus. This is analogous to the maximization step of the expectation–maximization (EM) algorithm. In our case, the expectation step is computing the predicate/argument alignment probabilities.

4 Experiment

4.1 Setup

We used a portion of OntoNotes Release 5.0[1] (with additional nominal/adjective predicates) that has Chinese-English word alignment annotation[2] as the basis for evaluating semantic alignment. This composes around 2000 Xinhua News and 3000 broadcast conversation (CCTV and Phoenix) sentence pairs. Merging the 2 resources result in parallel sentences with gold Treebank, gold PropBank, and gold word alignment annotations, which we dub the triple-gold corpus.

To generate reference predicate-argument alignments, we ran the alignment system with a cutoff threshold of $F_{c,e} < 0.4$ (i.e., alignments with F-score below 0.4 are discarded) using all gold annotations. We selected a small random sample of the Xinhua output and found the output to have both high precision and recall, with only occasional discrepancies caused by possible word alignment errors (and was no worse than inter-annotator disagreements). For predicate-argument alignments using automatic word alignment input, we chose a cutoff threshold of $F_{c,e} < 0.2$.

We trained our Chinese SRL system (Wu and Palmer, 2015) with Berkeley Parser output on Chinese PropBank 1.0 (all Xinhua News, excluding files in the triple-gold corpus). We trained our English SRL system (same architecture as the Chinese SRL system) with Berkley parser output on OntoNotes Release 5.0 (excluding files in the triple-gold corpus) and BOLT phase 1 data (which also includes nominal annotation). We use the Berkeley aligner trained on a 1.6M sentence parallel corpora collected from a variety of sources[3]. These same corpora were also used to build our probabilistic alignment model.

4.2 Alignment with Nominal/Adjective Predicates

We evaluated the impact of alignment with the addition of non-verb predicates on Xinhua News, as the broadcast conversation sections lack Chinese nominal annotations. In table 3, we restrict alignments to between only verb predicates, verb predicates with the addition of Chinese nominal predicates, and

[1] LDC2013T19

[2] LDC2009E83

[3] LDC2002E18, LDC2002L27, LDC2003E07, LDC2003E14, LDC2004T08, LDC2005E83, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006E24, LDC2006E26, LDC2006E34, LDC2006E85, LDC2006E86, LDC2006E92, LDC2006E93

pred. type	V_c-V_e	N_c-V_e	V_c-N_e	N_c-N_e	total
verb only	4879	-	-	-	4879
+Ch. nom.	4762	274	-	-	5036
+En. nom.	4849	-	384	-	5233
all pred.	4759	239	314	760	6072

Table 2: Predicate-argument mapping counts on Xinhua News, where only verb predicate annotations were available or verb and nominal/adjective predicate annotations. V_c represents Chinese verb predicates, N_e represents English nominal/adjective predicates

verb predicates with the addition of English nominal/adjective predicates, as well as allowing all predicate types.

The results show that the addition of nominal and adjective predicates for both English and Chinese increased the overall number of aligned Chinese and English predicate-argument structures by 24.5%. While a large portion of the additional alignments are of the non-verb to non-verb types, the availability of the non-verb predicates also allowed some previously unaligned verb predicates to align to non-verb predicates. This increased the total number of aligned Chinese verb predicates by 4.0% and aligned English verb predicates by 2.4%. Also, some verb predicates that were previously forced to align to another verb predicate have now found a more semantically similar non-verb predicate (evident by the decreased overall number of verb-to-verb alignments).

4.3 Alignment Probability Model

We produced the alignment probability model using the 1.6M sentence pair corpus, The EM algorithm converged after 2-3 iterations, as the alignments did not vary wildly with different α and β values (optimal $\alpha = 0.15$, $\beta = 0.1$). In general, the choice of β had a smaller impact on the overall mapping score of the corpus than α.

The results, detailed in table 3, show that using automatic SRL and word alignment, the probability model improved semantic alignment by about 1 F point on both Xinhua News (includes non-verb predicates) and broadcast conversation (verb predicates only for Chinese) sections. These improvements were found to be statistically significant[4] ($p \leq 0.01$). Surprisingly, the probability model (which was extracted from automatic SRL output), was able to improve the performance of the system using gold standard SRL input by 0.78 F point on broadcast conversation (also statistically significant w/ $p \leq 0.01$). For Xinhua News, the already very high baseline (92.40 F1) likely prevented any additional improvements.

With gold word alignment input, however, the probability model was not able to improve the results of either corpus section, even though the performances are lower than when using gold SRL inputs. This is not surprising as the probability model can suggest more semantically coherent alignments when faced with word alignment errors, but does not actually correct any input SRL mistakes made by automatic systems.

We also experimented with building the probability model using only 10% of the data. The improvements were generally 0.1-0.3 F points less than using the full dataset. The optimal $\alpha = 0.15$ and $\beta = 0.1$ did not change.

Inspecting the output, we found the probabilistic alignment system was able to correct the bad alignment example in figure 2 (corrected in figure 3), as the aligner preferred the more probable `ARG1` to `ARG1` alignment between 自筹 and *use* instead of the less probable `ARG1` to `ARGM-MNR` alignment between 自筹 and *build*. This also allowed the correct alignment between 建设/*construct* and *build* (also boosted by the increased predicate-to-predicate alignment probability).

While the predicate alignment performance difference between using automatic SRL and gold standard SRL input is around 7 F points, there is a much larger gap in core argument alignment performance: on Xinhua News, automatic SRL based output produced a 73.83 F-score While this is comparable to Fung et al. (Fung et al., 2007)'s 72.5 (albeit with different sections of the corpus and based on gold standard predicates from a bi-lingual dictionary), it's 18.27 F points lower than using gold standard SRL based output. When including all arguments, automatic SRL based output achieved 69.14% while the gold SRL based output achieved 87.56%. The performance on broadcast conversation shows a similar

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

corpus	system	predicate pair			core argument label			all argument label		
		p	r	f1	p	r	f1	p	r	f1
Xinhua News	baseline	86.93	82.56	84.69	80.27	67.04	73.06	75.14	62.53	68.26
	+prob model	87.97	83.47	**85.66**	81.07	67.78	**73.83**	76.64	62.98	**69.14**
	gold SRL	93.67	91.16	92.40	94.45	89.93	92.13	90.71	84.63	87.56
	+prob model	93.02	90.38	91.68	93.91	89.54	91.67	90.62	83.26	86.78
	gold WA	90.83	87.42	89.09	83.16	71.55	76.92	80.42	71.11	75.48
	+prob model	91.21	87.45	89.29	83.48	71.57	77.07	80.84	70.64	75.40
broadcast conversation	baseline	80.45	78.50	79.46	72.87	57.77	64.45	64.88	51.89	57.66
	+prob model	81.52	79.51	**80.50**	73.75	58.40	**65.18**	66.28	52.27	**58.45**
	gold SRL	89.50	85.29	87.34	90.21	82.19	86.02	82.61	75.20	78.73
	+prob model	90.17	86.15	**88.12**	90.82	82.93	**86.70**	84.11	75.25	**79.43**
	gold WA	87.02	86.66	86.84	78.31	65.11	71.10	74.85	64.94	69.55
	+prob model	87.17	86.61	86.89	78.26	64.80	70.89	74.96	64.20	69.16

Table 3: Predicate-argument mapping improvements using the probability model

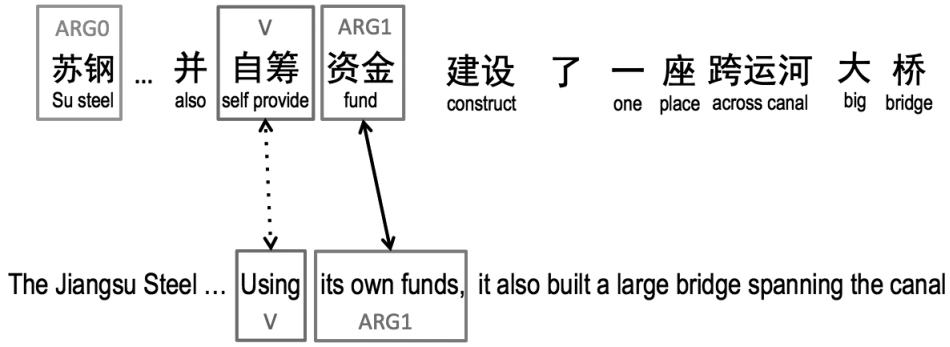

Figure 3: Corrected alignment using the probability model

drop between the 2 SRL outputs. Still, the probability model was able to generate statistically significant improvements to argument alignments when using automatic SRL inputs, albeit with a smaller margin.

These argument results are not too surprising given the alignment system need to deal with many sources of error, from errors introduced by the automatic Chinese SRL, English SRL and word alignment systems to incompatibilities between English and Chinese frame files, as well as confusions arising from implicit arguments. Along with the lack of improvement in predicate alignment performance when the probability alignment model uses gold word alignment input, the results indicate that a higher-performing PropBank alignment system need to address automatic SRL errors.

5 Conclusion

We described 2 improvements to Chinese-English PropBank alignments. The first takes advantage of expanded English nominal/adjective predicate annotation to produce a more comprehensive PropBank alignment between Chinese and English, increasing the number of aligned Chinese and English predicate-argument structures by 24.5%. The second utilizes predicate-argument alignment probabilities extracted from a large unannotated parallel corpus to both improve predicate-argument alignment performance and provide a probability model that can be used to evaluate/improve semantically-driven machine translation.

Given that the probability model, built using all automatic system output, provides smaller improvements to (or even degrades) the system when either gold standard SRL or word alignment is used, it still

has room for improvement. One such possible improvement would be to build a probability model predicated on verb classes/clusters. This could address the sparse alignment frequency count issue from the many possible Chinese-English predicate-argument pairings. For English, we can use the existing VerbNet class resource and train an automatic system for polysemous verbs. For Chinese, however, we would need to either induce verb classes through mapping (Wu et al., 2010), or via an automatic verb clustering method.

While we have achieved good predicate-argument alignment performance, specific argument alignment performance still lags behind. One reason is that while we can induce correct predicate-argument mapping from the argument mapping pairs, even when the predicates themselves are misaligned, for argument alignment, our system currently does not attempt to directly correct argument labels from automatic SRL output. Therefore, any SRL labeling error in the automatic SRL system output (made worse by having 2 languages) is propagated through the alignment system. A joint-inference/joint-learning framework between semantic alignment, SRL (including joint inference of Chinese and English SRL as proposed by Zhuang and Zong (2010)), and word alignment could potentially address the shortcomings in our current implementation.

Acknowledgement

We gratefully acknowledge the support of the National Science Foundation CISE-IISRI-0910992, Richer Representations for Machine Translation and, DARPA FA8750-09-C-0179 (via BBN) Machine Reading: Ontology Induction: Semlink+. 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

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. 2014. Propbank: Semantics of new predicate types. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jinho D. Choi, Martha Palmer, and Nianwen Xue. 2009. Using parallel propbanks to enhance word-alignments. In *Proceedings of ACL-IJCNLP workshop on Linguistic Annotation (LAW'09)*, pages 121–124.

Pascale Fung, Zhaojun Wu, Yongsheng Yang, and Dekai Wu. 2007. Learning bilingual semantic frames: Shallow semantic parsing vs. semantic role projection. In *11th Conference on Theoretical and Methodological Issues in Machine Translation*, pages 75–84.

William A. Gale. 1995. Good-turing smoothing without tears. *Journal of Quantitative Linguistics*, 2.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'07), demonstration session*, pages 177–180.

Chi-kiu Lo and Dekai Wu. 2011. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In *In Proceedings of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*.

Chi-kiu Lo, Karteek Addanki, Markus Saers, and Dekai Wu. 2013. Improving machine translation by training against an automatic semantic frame based evaluation metric. In *In Proceedings of 51th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2013)*.

Wei-Yun Ma. 2014. *Hybrid System Combination for Machine Translation: An Integration of Phrase-level and Sentences-level Combination Approaches*. Ph.D. thesis, Columbia University.

Nitin Madnani, Philip Resnik, Bonnie Dorr, and Richard Schwartz. 2008a. Applying automatically generated semantic knowledge: A case study in machine translation. In *NSF Symposium on Semantic Knowledge Discovery, Organization and Use*.

Nitin Madnani, Philip Resnik, Bonnie Dorr, and Richard Schwartz. 2008b. Are multiple reference translations necessary? investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA'08)*.

David Mareček. 2009. Using tectogrammatical alignment in phrase-based machine translation. In *Proceedings of WDS 2009 Contributed Papers*, pages 22—27.

Sebastian Padó and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 859–866, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 1161–1168, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, pages 71–106.

Philip Resnik. 2004. Exploiting hidden meanings: Using bilingual text for monolingual annotation. In Alexander Gelbukh, editor, *Lecture Notes in Computer Science 2945: Computational Linguistics and Intelligent Text Processing*, pages 283–299. Springer.

Dekai Wu and Pascale Fung. 2009a. Can semantic role labeling improve smt? In *Proceedings of the 13th Annual Conference of the EAMT*, pages 218–225, Barcelona, Spain.

Dekai Wu and Pascale Fung. 2009b. Semantic roles for smt: A hybrid two-pass model. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT'09)*, pages 13–16.

Shumin Wu and Martha Palmer. 2011. Semantic mapping using automatic word alignment and semantic role labeling. In *Proceedings of ACL-HLT workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5)*.

Shumin Wu and Martha Palmer. 2015. Can selectional preference help automatic semantic role labeling? In *Proceedings of the 4th Joint Conference on Lexical and Computational Semantics (⋆SEM)*.

Shumin Wu, Jinho D. Choi, and Martha Palmer. 2010. Detecting cross-lingual semantic similarity using parallel propbanks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*.

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.

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.