Semantic Mapping Using Automatic Word Alignment and Semantic Role Labeling

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

Martha Palmer
Department of Linguistics
Univerisity of Colorado at Boulder
martha.palmer@colorado.edu

Abstract

To facilitate the application of semantics in statistical machine translation, we propose a broad-coverage predicate-argument structure mapping technique using automated resources. Our approach utilizes automatic syntactic and semantic parsers to generate Chinese-English predicate-argument structures. The system produced a many-to-many argument mapping for all PropBank argument types by computing argument similarity based on automatic word alignment, achieving 80.5% F-score on numbered argument mapping and 64.6% F-score on all arguments. By measuring predicate-argument structure similarity based on the argument mapping, and formulating the predicate-argument structure mapping problem as a linear-assignment problem, the system achieved 84.9% F-score using automatic SRL, only 3.7% F-score lower than using gold standard SRL. The mapping output covered 49.6% of the annotated Chinese predicates (which contains predicate-adjectives that often have no parallel annotations in English) and 80.7% of annotated English predicates, suggesting its potential as a valuable resource for improving word alignment and reranking MT output.

1 Introduction

As the demand for semantically consistent machine translation rises (Wu and Fung, 2009a), the need for a comprehensive semantic mapping tool has become more apparent. With the current architecture of machine translation decoders, few ways of incorporating semantics in MT output include using word sense disambiguation to select the correct target translation (Carpuat and Wu, 2007) and reordering/reranking MT output based on semantic consistencies (Wu and Fung, 2009b) (Carpuat et al., 2010). While a comprehensive semantic mapping tool can supplement or improve the results of such techniques, there are many other exciting ideas we can explore: with automatic SRL, we can improve coverage (and possibly accuracy) of Chinese semantic class generation (Wu et al., 2010) by running the system on a large, unannotated parallel corpus. Using predicate-argument mappings as constraints, it may be possibly to improve SRL output by performing joint inference of SRL in source and target languages simultaneously, much like what Burkett and Klein (2008) was able to achieve with syntactic parsing.

As the foundation of many machine translation decoders (DeNeefe and Knight, 2009), word alignment has continuously played an important role in machine translation. There have been several attempts to improve word alignment, most of which have focused on tree-to-tree alignments of syntactic structures (Zhang et al., 2007; Mareček, 2009a). Our hypothesis is that the predicate-argument structure alignments can abstract away from language specific syntactic variation and provide a more robust, semantically coherent alignment across sentences.

We begin by running GIZA++ (Och and Ney, 2003), one of the most popular alignment tools, to obtain automatic word alignments between parallel English/Chinese corpora. To achieve a broader coverage of semantic mappings than just those anno-

Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 21–30,
ACL HLT 2011, Portland, Oregon, USA, June 2011.

tated in parallel PropBank-ed corpora, we attempt to map automatically generated predicate-argument structures. For each Chinese and English verb predicate pairs within a parallel sentence, we examine the quality of both the predicate and argument alignment (using GIZA++ word alignment output) and devise a many-to-many argument mapping technique. From that, we pose predicate-argument mapping as a linear assignment problem (optimizing the total similarity of the mapping) and solve it with the Kuhn-Munkres method (Kuhn, 1955). With this approach, we were able to incur only a small predicate-argument F-score degradation over using manual PropBank annotation. The output also provides much more fine-grained argument mapping that can be used for downstream MT applications.

2 Related work

Our basic approach to semantic mapping is similar to the idea of semantic similarity based on triangulation between parallel corpora outlined in Resnik (2004) and Madnani et al. (2008a; 2008b), but is implemented here quite differently. It is most similar in execution to the work of (Mareček, 2009b), which improves word alignment by aligning tectogrammatical trees in a parallel English/Czech corpus. The Czech corpus is first lemmatized because of the rich morphology, and then the word alignment is "symmetrized". However, this approach does not explicitly make use of the predicate-argument structure to confirm the alignments or to suggest new ones.

Padó and Lapata (2005; 2006) used word alignment and syntax based argument similarity to project English FrameNet semantic roles to German. The approach relied on annotated semantic roles on the source side only, precluding joint inferenece of the projection using reference or automatic target side semantic roles.

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 arguments and tuning on an unannotated parallel corpus. The system differs from ours in that it only provided one-to-one mapping of numbered arguments and may not be able to detect predicate mapping with no lexical relations that are nevertheless semantically related. Later, Wu and Fung (2009b) used parallel semantic roles to improve MT system outputs. Given the outputs from Moses (Koehn et al., 2007), a machine translation decoder, they reordered the outputs based on the best predicate-argument mapping. 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.

Choi et al. (2009) (and later Wu et al. (2010)) showed how to enhance Chinese-English verb alignments by exploring predicate-argument structure alignment using parallel PropBanks. The resulting system showed improvement over pure GIZA++ alignment. Those two systems differs from ours in that they operated on gold standard parses and semantic roles. The systems also did not provide explicit argument mapping between the aligned predicate-argument structures.

3 Resources

To perform automatic semantic mapping, we need an annotated corpus to evaluate the results. In addition, we also need a word aligner, a syntactic parser, and a semantic role labeler (as well as annotated and unannotated corpora to train each system).

3.1 Corpus

We used the portion of the Penn Chinese TreeBank with word alignment annotation as the basis for evaluating semantic mapping. The word-aligned portion, containing around 2000 parallel sentences, is exclusive to Xinhua News (and covers around 50% of the Xinhua corpus in the Chinese TreeBank). We then merged the word alignment annotation with the TreeBank and PropBank annotation of Ontonotes 4.0 (Hovy et al., 2006), which includes a wide array of data sources like broadcast news, news wire, magazine, web text, etc. A small percentage of the 2000 sentences were discarded because of tokenization differences. We dubbed the resulting 1939 parallel sentences as the triple-gold Xinhua corpus.

3.2 Word Alignment

We chose GIZA++ (Och and Ney, 2003) as our word alignment tool primarily because of its popularity, though there are other alternatives like Lacoste-Julien et al. (2006).

3.3 Phrase Structure Parsing

We chose the Berkeley Parser (Petrov and Klein, 2007) for phrase structure parsing since it has been tested on both English and Chinese corpora and can be easily retrained.

3.4 Semantic Role Labeling

For semantic role labeling (SRL), we built our own system using a fairly standard approach: SRL is posed as a multi-class classification problem requiring the identification of argument candidates for each predicate and their argument types. Typically, argument identification and argument labeling are performed in two separate stages because of time/resource constraints during training/labeling. For our system, we chose LIBLINEAR (Fan et al., 2008), a library for large linear classification problems, as the classifier. This alleviated the need to separate the identification and labeling stages: argument identification is trained simply by incorporating the "`NOT-ARG`" label into the training data.

Most the of the features used by the classifier are standard features found in many SRL systems; these include:

Predicate predicate lemma and its POS tag

Voice indicates the voice of the predicate. For English, we used the six heuristics detailed by Igo (2007), which detects both ordinary and reduced passive constructions. For Chinese, we simply detected the presence of passive indicator words (those with `SB`, `LB` POS tags) amongst the siblings of the predicate.

Phrase type phrase type of the constituent

Subcategorization phrase structure rule expanding the predicate parent

Head word the head word and its POS tag of the constituent

Parent head word whether the head word of the parent is the same as the head word of the constituent

Position whether the constituent is before or after the predicate

Path the syntactic tree path from the predicate to the constituent (as well as various path generalization methods)

First word first word and its POS tag of the constituent

Last word last word and its POS tag of the constituent

Syntactic frame the siblings of the constituent

Constituent distance the number of potential constituents with the same phrase type between the predicate and the constituent

We also created many bigrams (and a few trigrams) of the above features.

By default, LIBLINEAR uses the one-vs-all approach for multi-class classification. This does not always perform well for some easily confusable class labels. Also, as noted by Xue (2004), certain features are strong discriminators for argument identification but not for argument labeling, while the reverse is true for others. Under such conditions, mixing arguments and non-arguments within the same class may produce sub-optimal results for a binary classifier. To address these issues, we built a pairwise multi-class classifier (using simple majority voting) on top of LIBLINEAR.

The resulting English SRL system, evaluated using the CoNLL 2005 methodology, achieved a 77.3% F-score on the WSJ corpus, comparable to the leading system (Surdeanu and Turmo, 2005) using a single parser output. The Chinese SRL system, on the other hand, achieved 74.4% F-score on the triple-gold Xinhua corpus (similar but not directly comparable to Wu et al. (2006) and Xue (2008) because of differences in TreeBank/PropBank revisions as well as differences in test set).

4 Predicate-arguments mapping

4.1 Argument mapping

To produce a good predicate-argument mapping, we needed to consider 2 things: whether good argument mapping can be produced based on argument type only, and whether each argument only maps to one argument in the target language.

4.1.1 Predicate-dependent argument mapping

Theoretically, PropBank numbered arguments are supposed to be consistent across predicates: ARG0 typically denotes the agent of the predicate and ARG1 the theme. While this consistency may hold true for predicates in the same language, as Fung et al. (2007) noted, this is not a reliable indicator when mapping predicate-arguments between Chinese and English. For example, when comparing the PropBank frames of the English verb *arrive* and the synonymous Chinese verb 抵达, we see ARG1 (entity in motion) for *arrive*.01 is equivalent to ARG0 (agent) of 抵达.01 while ARG4 (end point, destination) is equivalent to ARG1 (destiny).

4.1.2 Many-to-many argument mapping

Just as there are shortcomings in assuming predicate independent argument mappings, assuming one-to-one argument mapping may also be overly restrictive. For example, in the following Chinese sentence:

大 通道 建设 **搞活** 了 大 西南 的 物流

big passage construction invigorated big southwest's material flow

the predicate 搞活(invigorate) has 2 arguments:

- ARG0: 大 通道 建设 (big passage construction)
- ARG1: 大 西南 的 物流 (big southwest's material flow)

In the parallel English sentence:

*Construction of the main passage has **activated** the flow of materials in the great southwest*

activate has 3 arguments:

- ARG0: *construction of the main passage*
- ARG1: *the flow of materials*
- ARGM-LOC: *in the great southwest*

In these parallel sentences, ARG1 of 搞活 should be mapped to both ARG1 and ARGM-LOC of *activate*.

While the English translation of 搞活, *invigorate*, is not a direct synonym of *activate*, they at least have some distant relationship as indicated by sharing the inherited hypernym *make* in the WordNet (Fellbaum, 1998) database. The same cannot be said for all predicate-pairs. For example, in the following parallel sentence fragments:

街 上 客流 如 潮

*on the street people **flow** like the tide*

the Chinese predicate-argument structure for 如(like) is:

- ARG0: 客流 (flow of guests)
- ARG1: 潮 (tide)
- ARGM-LOC: 街 上 (on the street)

while the English predicate-argument structure for *flow* is:

- ARG1: *people*
- ARGM-LOC: *on the street*
- ARGM-MNR: *like the tide*

Semantically, the predicate-argument pairs are equivalent. The argument mapping, however, is more complex:

- 如.ARG0 $\Longleftrightarrow$ *flow*.ARG1, *flow*.V
- 如.V, 如.ARG1 $\Longleftrightarrow$ *flow*.ARGM-MNR
- 如.ARGM-LOC $\Longleftrightarrow$ *flow*.ARGM-LOC

Table 1 details the argument mapping for the triple-gold Xinhua data. The mapping distribution for ARG0 and ARG1 is relatively deterministic (and similar to ones found by Fung et al. (2007)). Mappings involving ARG2-5 and modifier arguments, on the other hand, are much more varied. Typically, when there is a many-to-many argument mapping, it's constrained to a one-to-two or two-to-one mapping. Much more rarely is there a case of a two-to-two or even more complex mapping.

4.2 Word alignment based argument mapping

To achieve optimal mappings between parallel predicate-argument structure, we would like to maximize the number of words in the mapped argument set (over the entire set of arguments) while minimizing the number of unaligned words in the mapped argument set.

Let $a_{c,i}$ and $a_{c,j}$ denote arguments in Chinese and English respectively, A_I as a set of arguments, $W_{c,i}$ as words in argument $a_{c,i}$, and $map_e(a_i) = W_{e,i}$ as the word alignment function that takes the source argument and produces a set of words in the target

arg type	A0	A1	A2	A3	A4	ADV	BNF	DIR	DIS	EXT	LOC	MNR	PRP	TMP	TPC	V
A0	***1610***	79	25	0	0	28	1	0	0	0	8	5	1	11	1	9
A1	432	***2665***	128	**11**	0	83	**9**	**12**	0	0	29	12	5	21	**3**	142
A2	43	*310*	**140**	8	3	55	6	9	0	**2**	20	10	1	4	1	67
A3	2	14	*21*	7	0	2	4	2	0	0	1	2	1	0	1	4
A4	1	*37*	9	3	**6**	0	0	0	0	0	1	0	1	0	0	4
ADV	33	36	9	6	0	***307***	2	5	6	0	44	**121**	6	11	2	19
CAU	1	0	0	0	0	1	0	0	0	0	0	0	*16*	0	0	1
DIR	1	13	3	2	0	1	0	3	0	0	3	0	0	0	0	*20*
DIS	2	0	0	0	0	*69*	0	0	**40**	0	2	1	3	3	0	0
EXT	0	4	0	0	0	*26*	0	0	0	0	0	0	0	0	0	2
LOC	23	*65*	13	1	0	3	1	0	0	0	**162**	0	0	5	0	4
MNR	9	9	5	0	0	*260*	0	0	0	1	3	34	0	0	0	25
MOD	1	0	0	0	0	*159*	0	0	0	0	0	0	0	0	0	84
NEG	0	0	0	0	0	*24*	0	0	0	0	0	0	0	0	0	5
PNC	3	23	11	4	0	1	6	1	0	0	1	2	***35***	2	0	8
PRD	0	0	*3*	0	0	0	0	0	0	0	0	0	0	0	0	1
TMP	14	21	2	0	0	235	0	3	0	1	8	16	0	***647***	0	6
V	25	28	22	1	0	211	1	0	1	0	2	12	0	0	0	***3278***

Table 1: Chinese argument type (column) to English argument type (row) mapping on triple-gold Xinhua corpus

language sentence. We define precision as the fraction of aligned target words in the mapped argument set:

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

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

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

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

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

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

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

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

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

To generate the set of argument mapping pairs, we simply choose all pairs of $a_{c,i}, a_{e,j} \in A_{c,I}, A_{e,J}$ where $F_{ij} \geq \epsilon$ ($\epsilon > 0$).

Directly optimizing equation 4 requires exhaustive search of all argument set combinations between the source and target, which is NP-complete. While the typical number of arguments for each predicate is relatively small, this is nevertheless inefficient. We performed the following greedy-based approximation with quadratic complexity:

1. Compute the best (based on F-score of equation 5) pair of source-target argument mappings for each source argument (target argument may be reused)
2. Select the remaining argument pair with the highest F-score
3. Insert the pair in $A_{c,I}, A_{e,J}$ if it increases F_{IJ}, else discard
4. repeat until all argument pairs are exhausted
5. repeat 1-4 reversing the source and target direction
6. merge the output of the 2 directions

Much like GIZA++ word alignment where the output of each direction produces only one-to-many mappings, merging the output of the two directions produces many-to-many mappings.

4.3 One-to-one predicate-argument mapping

To find the best predicate-argument mapping between Chinese and English parallel sentences, we assume each predicate in a Chinese or English sentence can only map to one predicate in the target sentence. As noted by Wu et al. (2010), this assumption is mostly valid for the Xinhua news corpus, though occasionally, a predicate from one sentence may align more naturally to two predicates in the target sentence. This typically occurs with verb conjunctions. For example the Chinese phrase "观光 旅游" (sightseeing and tour) is often translated to the single English verb "travel". As noted by Xue and Palmer (2009), the Chinese PropBank annotates predicative adjectives, which tend not to have an equivalent in the English PropBank. Additionally, some verbs in one language are nominalized in the other. This results in a good portion of Chinese or English predicates in parallel sentences not having an equivalent in the other language.

With the one-to-one mapping constraint, we optimize the mapping by maximizing the sum of the F1-scores (as defined by equation 4) of the predicates and arguments in the mapping. Let P_C and P_E denote the sets of predicates in Chinese and English respectively, with $G(P_C, P_E) = \{g : P_C \mapsto P_E\}$ as the set of possible mappings between the two predicate sets, then the optimal mapping is:

$$g^* = \arg\max_{g \in G} \sum_{i,j \in g} F_{C_i,E_j} \quad (6)$$

To turn this into a classic linear assignment problem, we define $Cost(P_{C_i}, P_{E_j}) = 1 - F_{C_i,E_j}$, and (6) becomes:

$$g^* = \arg\min_{g \in G} \sum_{i,j \in g} Cost(P_{C,i}, P_{E,j}) \quad (7)$$

(7) can be solved in polynomial time with the *Kuhn-Munkres* algorithm (Kuhn (1955)).

5 Experimental setup

5.1 Reference predicate-argument mapping

To generate reference predicate-argument mappings, we ran the mapping system described in section 4.2 with a cutoff threshold of $F_{C_i,E_j} < 0.65$ (i.e., alignments with F-score below 0.65 are discarded). We reviewed a small random sample of the output and found it to have both high precision and recall, with only occasional discrepancies caused by possible word alignment errors. If one-to-one argument mapping is imposed, the reference predicate-argument mapping will lose 8.2% of the alignments. For mappings using automatic word alignment, we chose a cutoff threshold of $F_{C_i,E_j} < 0.15$. This can easily be tuned for higher precision or recall based on application needs.

5.2 Parser, SRL, GIZA++

We trained the Berkeley parser and our SRL system on Ontonotes 4.0, excluding the triple-gold Xinhua sections as well as the non-English or Chinese sourced portion of the corpus. GIZA++ was trained on 400K parallel Chinese-English sentences from various sources with the default parameters. For the word mapping functions $map_e(a_c)$, $map_c(a_e)$ in equation 5, instead of taking the word alignment intersection of the source-target and target-source directions as Padó and Lapata (2006), we used the two alignment outputs seperately (using the Chinese-English output when projecting Chinese argument to English words, and vice versa). On average (from the 400K corpus), an English sentence contains 28.5% more tokens than the parallel Chinese sentence (even greater at 36.2% for the Xinhua portion). Taking either the intersection or union will significantly affect recall or precision of the alignment.

6 Results

6.1 Semantic role labeling

We first provide some results of the SRL system on the triple-gold Xinhua corpus in table 2. Unlike the conventional wisdom which expects English SRL to outperform Chinese SRL, when running on the Chinese-sourced Xinhua parallel corpus, our SRL actually performed better on Chinese than English (74.4% vs 71.8% F-score). The Berkeley parser output also seemed to be of higher quality on Chinese; the system was able to pick out better constituent candidates in Chinese than English, as evidenced by the higher recall for oracle SRL (92.6% vs 91.1%). Comparing the quality of the output by argument type, we found the only argument type where the Chinese SRL system performed signifi-

language	type	P	R	F1
Chinese	CoNLL	77.9%	71.1%	**74.4%**
	oracle	100%	92.6%	96.1%
	word match	84.8%	74.6%	79.4%
English	CoNLL	75.6%	68.4%	**71.8%**
	oracle	100%	91.1%	95.2%
	word match	82.7%	69.4%	75.5%

Table 2: SRL results on triple-gold Xinhua corpus. "arg match" is the standard CoNLL 2005 evaluation metric, "oracle" is the oracle SRL based on automatic parser output, and "word match" is scoring based on length of argument overlap with the reference

Evaluation	gold	P	R	F1
predicate-argument	yes	88.7%	88.5%	88.6%
	no	84.6%	85.3%	**84.9%**
A0-5 label	yes	97.8%	96.2%	97.0%
	no	87.0%	74.9%	**80.5%**
A0-5 span	no	67.9%	57.9%	62.5%
all arg label	yes	84.0%	79.3%	81.6%
	no	70.3%	59.8%	*64.6%*
all arg span	no	61.6%	52.2%	56.5%

Table 3: Predicate-argument mapping results

cantly worse is ARG0 (almost 10% F-score lower). This is likely caused by dropped pronouns in Chinese sentences (Yang and Xue, 2010), making it harder for both the syntactic and semantic parsers to identify the correct subject.

We also report the SRL result scored at word level instead of at argument level (79.4% F-score for Chinese and 75.5% for English). The CoNLL 2005 shared task scoring (Surdeanu and Turmo, 2005) discounts arguments that are not a perfect word span match, even if the system output is semantically close to the reference argument. While this is important in some applications of SRL, for other applications like improving word alignment with SRL, improving recall on approximate arguments may be a better trade-off than having high precision on perfectly matched arguments. We noticed that while overall improvement in SRL improves both word level and argument level performance, for otherwisely identical systems, we can slightly favor word level performance (up to 1-3% F-score) by including positive training samples that are not a perfect argument match.

6.2 Predicate-argument mapping

Table 3 details the results of Chinese-English predicate-argument mapping. Using automatic SRL and word alignment, the system achieved an 84.9% F-score, only 3.7% F-score less than using gold standard SRL annotation. When looking at only arguments, however, the differences are larger: automatic SRL based output produced an 80.5% F-score for core arguments. While this compares favorably to Fung et al. (2007)'s 72.5% (albeit with different sections of the corpus), it's 16.5% F-score lower than gold SRL based output. When including all arguments, automatic SRL based output achieved 64.6% while the gold SRL based output achieved 81.6%. This indicates that the mapping result for all arguments is limited by errors in word alignment. We also report the results of automatic SRL on both producing the correct argument mappings and word spans (62.5% for core arguments and 56.5% for all arguments). This may be relevant for applications such as joint inference between word alignment and SRL.

We also experimented with discriminative (reweighing) word alignment based on part-of-speech tags of the words to improve the mapping system but were not able to achieve better results. This may be due to the top few POS types accounting for most of the words in a language, therefore it did not prove to be a strong discriminator.

6.3 Mapping coverage

Table 4 provides predicate and word coverage details of the predicate-argument mapping, another potentially relevant statistic for applications of predicate-argument mapping. High coverage of predicates and words in the mappings may provide more relevant constraints to help reorder MT output or rerank word alignment. We expect labeling English nominalized predicate-arguments will help increase both predicate and word coverage in the mapping output.

In order to build a comprehensive probability model of Chinese-English predicate-argument mapping, we applied the mapping technique on an unannotated 400K parallel sentence corpus. Automatic

output	type	language	coverage
triple-gold	predicate	Chinese	50.0%
	predicate	English	81.3%
	word	Chinese	66.0%
	word	English	64.2%
automatic	predicate	Chinese	49.6%
	predicate	English	80.7%
	word	Chinese	57.4%
	word	English	55.4%

Table 4: Predicate-argument mapping coverage. Predicate coverage denotes the number of mapped predicates over all predicates in the corpus, word coverage denotes the number of words in the mapped predicate-arguments over all words in the corpus

language	PropBank verb framesets	appeared in corpus	appeared in mapping
Chinese	16122	8591	7109
English	5473	3689	3121

Table 5: Frameset coverage on the 400K parallel sentence corpus

SRL found 1.6 million Chinese predicate instances and 1.3 million English predicate instances. The mapping system found around 700K predicate-pairs (with $F_{C,E} < 0.3$). Table 5 shows the number of unique verbs in the corpus and contained in the mapping results within the Chinese and English PropBank verb framesets. The corpus also included some verbs that do not appear in PropBank framesets.

7 Conclusion and future work

We proposed a broad-coverage predicate-argument mapping system using automatically generated word alignment and semantic role labeling. We also provided a competitive Chinese and English SRL system using a LIBLINEAR classifier and pairwise multi-class classification approach. By exploring predicate-argument structure, the mapping system is able to generate mappings between semantically similar predicate-argument structures containing non-synonymous predicates, achieving an 84.9% F-score, only 3.7% lower than the F-score of gold-standard SRL based mappings. Utilizing word alignment information, the system was able to provide detailed many-to-many argument mappings (occurs in 8.2% of the reference mappings) for core arguments and modifier arguments, achieving an 80.5% F-score for core arguments and 64.6% F-score for all arguments.

While our experiment with discriminative word alignment based on POS tags did not show improvement, there are other word grouping/weighing metrics like n-gram based clustering, verb classification, term frequency, that may be more appropriate for semantic mapping. With the advent of a predicate-argument annotation resource for nominalization, Ontonotes 5, we plan to update our SRL system to produce nominalized predicate-arguments. This would potentially increase the predicate-argument mapping coverage in the corpus as well as increasing the accuracy of mapping (by reducing the number of unmappable predicate-arguments), making the mapping more useful for downstream applications.

We are also experimenting with a probabilistic approach to predicate-argument mapping to improve the robustness of mapping against word alignment errors. Using the output of the current system on a large corpus, we can establish models for $p(pred_e|pred_c)$, $p(arg_e|pred_c, pred_e, arg_c)$ and refine them through iterations of expectation-maximization. If this approach shows promise, the next step would be to explore integrating the mapping model directly into GIZA++ for joint inference of word alignment and predicate-argument mapping. Other statistical translation specific applications we would like to explore include extensions of MT output reordering (Wu and Fung, 2009b) and reranking using predicate-argument mapping, as well as predicate-argument projection onto the target language as an evaluation metric for MT output.

Acknowledgement

We gratefully acknowledge the support of the National Science Foundation Grants CISE- CRI-0551615, and a grant from the Defense Advanced Research Projects Agency (DARPA/IPTO) under the GALE program, DARPA/CMO Contract No. HR0011-06-C-0022, subcontract from BBN, Inc. 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

David Burkett and Dan Klein. 2008. Two languages are better than one (for syntactic parsing). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 877–886, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Marine Carpuat, Yuval Marton, and Nizar Habash. 2010. Improving arabic-to-english statistical machine translation by reordering post-verbal subjects for alignment. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 178–183.

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.

Steve DeNeefe and Kevin Knight. 2009. Synchronous tree adjoining machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, volume 2, pages 727–736.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

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

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.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90 In *Proceedings of HLT-NAACL 2006*, pages 57–60.

Sean Paul Igo. 2007. Identifying reduced passive voice constructions in shallow parsing environments. Master's thesis, University of Utah.

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.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Simon Lacoste-Julien, Ben Taskar, Dan Klein, and Michael I. Jordan. 2006. Word alignment via quadratic assignment. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 112–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

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. 2009a. Improving word alignment using alignment of deep structures. In *Proceedings of the 12th International Conference on Text, Speech and Dialogue*, pages 56–63.

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

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *In HLT-NAACL '07*.

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.

Mihai Surdeanu and Jordi Turmo. 2005. Semantic role labeling using complete syntactic analysis. In *Proceedings of CoNLL-2005 shared task*, pages 221–224.

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.

Zhaojun Wu, Yongsheng Yang, and Pascale Fung. 2006. C-assert: Chinese shallow semantic parser. `http://hlt030.cse.ust.hk/research/c-assert/`.

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*.

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the chinese treebank. *Nat. Lang. Eng.*, 15(1):143–172.

Nianwen Xue. 2004. Calibrating features for semantic role labeling. In *Proceedings of EMNLP 2004*, pages 88–94.

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

Yaqin Yang and Nianwen Xue. 2010. Chasing the ghost: recovering empty categories in the chinese treebank. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1382–1390, Stroudsburg, PA, USA. Association for Computational Linguistics.

Min Zhang, Hongfei Jiang, Ai Ti Aw, Jun Sun, Sheng Li, and Chew Lim Tan. 2007. A tree-to-tree alignment-based model for statistical machine translation. In *Machine Translation Summit XI*.