Using Cross-Lingual Projections to Generate Semantic Role Labeled Corpus for Urdu - A Resource Poor Language

Smruthi Mukund
CEDAR
University at Buffalo
smukund@buffalo.edu

Debanjan Ghosh
Thomson Reuters R&D
debanjan.ghosh@
thomsonreuters.com

Rohini K. Srihari
CEDAR
University at Buffalo
rohini@cedar.buffalo.edu

Abstract

In this paper we explore the possibility of using cross lingual projections that help to automatically induce role-semantic annotations in the PropBank paradigm for Urdu, a resource poor language. This technique provides annotation projections based on word alignments. It is relatively inexpensive and has the potential to reduce human effort involved in creating semantic role resources. The projection model exploits lexical as well as syntactic information on an English-Urdu parallel corpus. We show that our method generates reasonably good annotations with an accuracy of 92% on short structured sentences. Using the automatically generated annotated corpus, we conduct preliminary experiments to create a semantic role labeler for Urdu. The results of the labeler though modest, are promising and indicate the potential of our technique to generate large scale annotations for Urdu.

1 Introduction

Semantic Roles (also known as thematic roles) help to understand the semantic structure of a document (Fillmore, 1968). At a fundamental level, they help to capture the similarities and differences in the meaning of verbs via the arguments they define by generalizing over surface syntactic configurations. In turn, these roles aid in domain independent understanding as the semantic frames and semantic understanding systems do not depend on the syntactic configuration for each new application domain. Identifying semantic roles benefit several language processing tasks - information extraction (Surdeanu *et al.*, 2003), text categorization (Moschitti, 2008) and finding relations in textual entailment (Burchardt and Frank 2006).

Automatically identifying semantic roles is often referred to as shallow semantic parsing (Gildea and Jurafsky, 2002). For English, this process is facilitated by the existence of two main SRL annotated corpora – FrameNet (Baker *et al.*, 1998) and PropBank (Palmer *et al.*, 2005). Both datasets mark almost all surface realizations of semantic roles. FrameNet has 800 semantic frames that cover 120,000 example sentences[1]. PropBank has annotations that cover over 113,000 predicate-argument structures. Clearly English is well supported with resources for semantic roles. However, there are other widely spoken resource poor languages that are not as privileged. The PropBank based resources available for languages like Chinese (Xue and Palmer, 2009), Korean (Palmer *et al.*, 2006) and Spanish (Taule, 2008) are only about two-thirds the size of the English PropBank.

Several alternative techniques have been explored in the literature to generate semantic role labeled corpora for resource poor languages as providing manually annotated data is time consuming and involves intense human labor. Ambati and Chen (2007) have conducted an extensive survey and outlined the benefits of using parallel corpora to transfer annotations. A wide range of annotations from part of speech (Hi and Hwa, 2005) and chunks (Yarowsky *et al.*, 2001) to word senses (Diab and Resnik, 2002), dependencies (Hwa *et al.*, 2002) and semantic roles (Pado and Lapata, 2009) have been successfully transferred between languages. FrameNet style annotations in Chinese is obtained by mapping English FrameNet entries directly to concepts listed in HowNet[2] (online ontology for Chinese) with an accuracy of 68% (Fung and Chen, 2004).

[1] Wikipedia - http://en.wikipedia.org/wiki/PropBank
[2] http://www.keenage.com/html/e_index.html

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 797–805,
Beijing, August 2010

Fung *et al.* (2007) analyze an automatically annotated English-Chinese parallel corpus and show high cross-lingual agreement for PropBank roles (range of 75%-95% based on the roles).

In this paper we explore the possibility of using English-Urdu parallel corpora to generate SRL annotations for Urdu, a less commonly taught language (LCTL). Earlier attempts to generate SRL corpora using annotation projections have been for languages such as German, French (Pado and Lapata, 2009) and Italian (Moschitti, 2009) that have high vocabulary overlap with English. Also, German belongs to the same language family as English (Germanic family). Urdu on the other hand is an Indic language that is grammatically very different and shares almost no vocabulary with English.

The technique of cross lingual projections warrants good BLEU score that ensures correct word alignments. According to NIST 2008 Open Machine Translation challenge [3], a 0.2280 best BLEU score was achieved for Urdu to English translation. This is comparable to the BLEU scores achieved for German to English – 0.253 and French to English – 0.3 (Koehn, 2005). But, for SRL transfer, perfect word alignment is not mandatory as SRL requires semantic correspondence only. According to Fillmore (1982) semantic frames are based on conceptual structures. They are generalizations over surface structures and hence less prone to syntactic variations. Since English and Urdu have a reasonable semantic correspondence (Example 3), we believe that the projections when capped with a post processing step will considerably reduce the noise induced by inaccurate alignments and produce acceptable mappings.

Hindi is syntactically similar to Urdu. These languages are standardized forms of Hindustani. They are free word order languages and follow a general SOV (Subject-Object-Verb) structure. Projection approach has been used by (Mukerjee *et al.*, 2006) and (Sinha, 2009) to transfer verb predicates from English onto Hindi. Sinha (2009) achieves a 90% F-Measure in verb predicate transfer from English to Hindi. This shows that using cross lingual transfer approach to obtain semantic annotations for Urdu from English is an idea worth exploring.

[3]http://www.itl.nist.gov/iaui/894.01/tests/mt/2008/doc/mt08_official_results_v0.html

1.1 Approach

Our approach leverages existing English PropBank annotations provided via the SemLink[4] corpus. SemLink provides annotations for VerbNet using the **pb (P**rop**B**ank) attribute. By using English-Urdu parallel corpus we acquire verb predicates and their arguments. When we transfer verb predicates (lemmas), we also transfer **pb** attributes. We obtain annotation projections from the parallel corpora as follows:

1. Take a pair of sentences *E* (in English) and *U* (in Urdu) that are translations of each other.
2. Annotate *E* with semantic roles.
3. Project the annotations from *E* onto *U* using word alignment information, lexical information and linguistic rules that involve syntactic information.

There are several challenges to the annotation projection technique. Dorr (1994) presents some major lexical-semantic divergence problems applicable in this scenario:

(a) Thematic Divergence - In some cases, although there exists semantic parallelism, the theme of the English sentence captured in the subject changes into an object in the Urdu sentence (Example 1).

(b) Conflatational Divergence - Sometimes target translations spans over a group of words (Example 1: *plays* is mapped to *kirdar ada)*. Trying to ascertain this word span for semantic roles is difficult as the alignments can be incomplete and very noisy.

(c) Demotional divergence and Structural divergence - Despite semantic relatedness, in some sentence pairs, alignments obtained from simple projections generate random matchings as the usage is syntactically dissimilar (Example 2).

Handling all challenges adds complexity to our model. The heuristic rules that we implement are guided by linguistic knowledge of Urdu. This increases the effectiveness of the alignments.

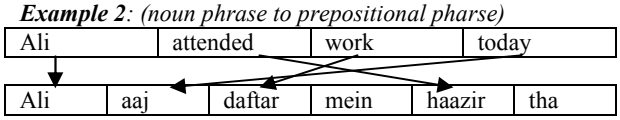
Example 1:

I *(subject)*	am	Angry	at	Reheem *(object)*
Raheem *(subject)*	mujhe *(object)*	Gussa	dilate	hai

(Raheem brings anger in me)

[4] http://verbs.colorado.edu/semlink/

***Example 2**: (noun phrase to prepositional pharse)*



(Ali was present at work today)

2 Generating Parallel Corpora

PropBank provides SRL annotated corpora for English. It uses predicate independent labels (ARG0, ARG1, etc.) which indicate how a verb relates to its arguments. The argument types are consistent across all uses of a single verb and do not consider the sense of the verb. We use the PropBank annotations provided for the Wall Street Journal (WSJ) part of the Penn Tree bank corpus (Marcus *et al.*, 2004). The arguments of a verb are labeled sequentially from ARG0 to ARG5 where ARG0 is the proto-typical Agent, ARG1 is the proto-typical patient, ARG2 is the recipient, and so on. There are other adjunct tags in the dataset that are indicated by ARGM that include tags for location (ARGM-LOC), temporal tags (ARGM-TMP) etc.

An Urdu corpus of 6000 sentences corresponding to 317 WSJ articles of Penn Tree Bank corpus is provided by CRULP[5] (used in the NIST 2008 machine translation task). We consider 2350 English sentences with PropBank annotations that have corresponding Urdu translations (CRULP corpus) for our experiments.

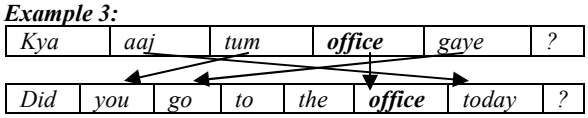
2.1 Sentence Alignment

Sentence alignment is a prerequisite for any parallel corpora processing. As the first step, we had to generate a perfect sentence aligned parallel corpus as the translated sentences, despite belonging to the same domain (WSJ – Penn tree bank), had several errors in demarcating the sentence boundaries.

Sentence alignment between English and Urdu is achieved over two iterations. In the first iteration, the length of each sentence is calculated based on the occurrence of words belonging to important part of speech categories such as proper nouns, adjectives and verbs. Considering main POS categories for length assessment helps overcome the conflatational divergence issue. For each English sentence, Urdu sentences with the same length are considered to be probable candidates for alignment. In the second iteration, an Urdu-English lexicon is used on the Urdu corpus and English translations are obtained. An English-Urdu sentence pair with maximum lexical match is considered to be sentence aligned.

Clearly this method is highly dependent on the existence of an exhaustive Urdu-English dictionary. The lexicons that we use to perform lookups are collected by mining Wikipedia and other online resources (Mukund *et al.,* 2010). However, lexicon lookups will fail for Out-Of-Vocabulary words. There could also be a collision if Urdu sentences have English transliterated words (Example 3, *"office"*). Such errors are manually verified for correctness.

Example 3:



2.2 Word Alignment

In the case of generating word alignments it is beneficial to calculate alignments in both translation directions (English – Urdu and Urdu - English). This nature of symmetry will help to reduce alignment errors. We use the Berkeley Aligner[6] word alignment package which implements a joint training model with posterior decoding (Liang *et al.*, 2006) to consider bidirectional alignments. Predictions are made based on the agreements obtained by two bidirectional models in the training phase. The intuitive objective function that incorporates data likelihood and a measure of agreement between the models is maximized using an EM-like algorithm. This alignment model is known to provide 29% reduction in AER over IBM model 4 predictions.

On our data set the word alignment accuracy is 71.3% (calculated over 200 sentence pairs). In order to augment the alignment accuracy, we added 3000 Urdu-English words and phrases obtained from the Urdu-English dictionary to our parallel corpus. The alignment accuracy improved by 3% as the lexicon affects the word co-occurrence count.

Word alignment in itself does not produce accurate semantic role projections from English to Urdu. This is because the verb predicates in Urdu can span more than one token. Semantic roles

[5]http://www.crulp.org/

[6] http://nlp.cs.berkeley.edu/Main.html

can cover sentential constituents of arbitrary length, and simply using word alignments for projection is likely to result in wrong role spans. Also, alignments are not obtained for all words. This could lead to missing projections.

One way to correct these alignment errors is to devise token based heuristic rules. This is not very beneficial as writing generic rules is difficult and different errors demand specific rules. We propose a method that considers POS, tense and chunk information along with word alignments to project annotations.

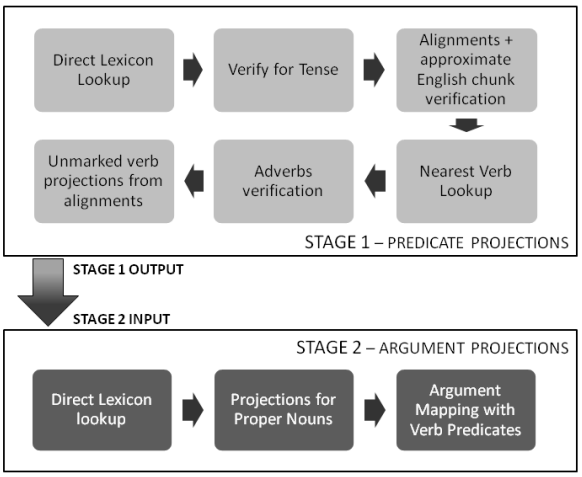


Figure 1: Projection model

Our proposed approach can be explained in two stages as shown in figure 1. In Stage 1 only verb predicates are transferred from English to Urdu. Stage 2 involves transfer of arguments and depends on the output of Stage 1. Predicate transfer cannot rely entirely on word alignments (§3). Rules devised around the chunk boundaries boost the verb predicate recognition rate.

Any verb group sequence consisting of a main verb and its auxiliaries are marked as a verb chunk. Urdu data is tagged using the chunk tag set proposed exclusively for Indian languages by Bharati *et al.*, (2006). Table 1 shows the tags that are important for this task.

Verb Chunk	Description
VGF	Verb group is finite (decided by the auxiliaries)
VGNF	Verb group for non-finite adverbial and adjectival chunk
VGNN	Verb group has a gerund

Table 1: Verb chunk tags in Urdu

The sentence aligned parallel corpora that we feed as input to our model is POS tagged for both English and Urdu. Urdu data is also tagged for chunk boundaries and morphological features like tense, gender and number information. Named Entities are also marked on the Urdu data set as they help in tagging the ARGM arguments. All the NLP taggers (POS, NE, Chunker, and Morphological Analyzer) used in this work are detailed in Mukund *et al.,* (2010).

English data is not chunked using a conventional chunk tagger. Each English sentence is split into virtual phrases at boundaries determined by the following parts of speech – IN, TO, MD, POS, CC, DT, SYM,: (Penn Tree Bank tagset). These tags represent positions in a sentence that typically mark context transitions (they are mostly the closed class words). We show later how these approximate chunks assist in correcting predicate mappings.

We use an Urdu-English dictionary (§2.1) that assigns English meanings to Urdu words in each sentence. Using translation information from a dictionary can help transfer verb predicates when the translation equivalent preserves the lexical meaning of the source language.

The first rule that gets applied for predicate transfer is based on lexicon lookup. If the English verb is found to be a synonym to an Urdu word that is part of a verb chunk, then the lemma associated with the English word is transferred to the entire verb chunk in Urdu. However not all translations' equivalents are lexically synonymous. Sometimes the word used in Urdu is different in meaning to that in English but relevant in the context (lexical divergence).

The word alignments considered in proximity to the approximate English chunks come to rescue in such scenarios. Here, for all the words occurring in each Urdu verb chunk, corresponding English aligned words are found from the word alignments. If the words that are found belong to the same approximate English chunk, then the verb predicate of that chunk (if present) is projected onto the verb chunk in Urdu. This heuristic technique increases the verb projection accuracy by about 15% as shown in §4.

The Penn tree bank tag set for English part of speech has different tags for verbs based on the tense information. VBD is used to indicate past tense, and VBP and VBZ for present tense. Urdu

also has the tense information associated with the verbs in some cases. We exploit this similarity to project the verb predicates from English onto Urdu.

The adverbial chunk in Urdu includes pure adverbial phrases. These chunks also form part of the verb predicates.

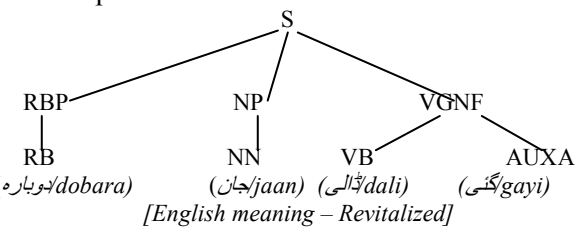


Figure 2: example for demotional divergence

E.g. consider the English word "revitalized" (figure 2). This is tagged VBD. However, the Urdu equivalent of this word is *"دوبارہ جان ڈالی گئی" (dobara jaan daali gayi ~ to put life in again)*. The POS tags are "RB, NN, VB, AUXA" *(adverb, noun, verb, aspectual auxiliary)*. The word *"dobara"* is a part of the adverbial chunk RBP and the infinite verb chunk VGNF spans across the last two words *"daali gayi"*. *"jaan"* is a noun chunk. This kind of demotional divergence is commonly observed in languages like Hindi and Urdu. In order to consider this entire phrase to be the Urdu equivalent representation of the English word *"revitalized"*, a rule for adverbial chunk is included as the last step to account for unaccommodated English verbs in the projections.

In the PropBank corpus, predicate argument relations are marked for almost all occurrences of non-copula verbs. We however do not have POS tags that help to identify non-copula words. Words that can be auxiliary verbs occur as non-copula verbs in Urdu. We maintain a list of such auxiliary verbs. When the verb chunk in Urdu contains only one word and belongs to the list, we simply ignore the verb chunk and proceed to the next chunk. This avoids several false positives in verb projections.

Stage 2 of the model includes the transfer of arguments. In order to see how well our method works, we project all argument annotations from English onto Urdu. We do not consider word alignments for arguments with proper nouns. The double metaphone algorithm (Philips 2000) is applied on both English NNP (proper noun) tagged words as well as English transliterated Urdu (NNP) tagged words. Arguments from English are mapped onto Urdu for word pairs with the same metaphone code.

For other arguments, we consider word alignments in proximity to verb predicates. The argument boundaries are determined based on chunk and POS information. We observe that our method projects the annotations associated with nouns fairly well. However, when the arguments contain adjectives, the boundaries are discontinuous. In such cases, we consider the entire chunk without the case marker as a probable candidate for the projected argument. We also have some issues with the ARGM-MOD arguments in that they overlap with the verb predicates. When the verb predicate that it overlaps with is a complex predicate, we consider the entire verb chunk to be the Urdu equivalent candidate argument. These rules along with word alignments yield fairly accurate projections.

The rules that we propose are dependent on the POS, chunk and tense information that are language specific. Hence our method is language independent only to the extent that the new language considered should have similar syntactic structure as Urdu. Indic languages fall in this category.

3 Verb Predicates

Detecting verb predicates can be a challenging task especially if very reliable and efficient tools such as POS tagger and chunkers are not available. We apply the POS tagger (CRULP tagset, 88% F-Score) and Chunker (Hindi tagset, 90% F-Score) provided by Mukund *et al.,* (2010) on the Urdu data set and show that syntactic information helps to compensate alignment errors. Stanford POS tagger[7] (Penn Tree bank tagset) is applied on the English data set.

Predicates can be simple predicates that lie within the chunk boundary or complex predicates when they span across chunk boundaries. When verbs in English are expressed in Urdu/Hindi, in several cases, more than one word is used to achieve perfect translation. In English the tense of the verb is mostly captured by the verb morpheme such as *"ask**ed**" "sa**id**" "say**ing**"*. In Urdu the tense is mostly captured by the auxiliary verbs. So a single word English verb such as *"talking"* would be translated into two words

[7] http://nlp.stanford.edu/software/tagger.shtml

"batein karna" where *"karna"~ do* is the auxiliary verb. However this cannot be generalized as there are instances when translations are word to word. E.g. *"said"* is mapped to a single word Urdu verb *"kaha"*.

Complex predicates in Urdu can occur in the following POS combinations. *Noun+Verb, Adjective+Verb, Verb+Verb, Adverb+Verb*. Table 2 lists the main verb tags present in the Urdu POS tagset. (refer Penn Tree bank POS tagset for English tags).

Urdu Tags	Description
VB	Verb
VBI	Infinitive Verb
VBL	Light Verb
VBLI	Infinitive Light Verb
VBT	Verb to be
AUXA	Aspectual Auxiliary
AUXT	Tense Auxiliary

Table 2: Verb tags

Auxiliary verbs in Urdu occur alongside VB, VBI, VBL or VBLI tags. Sinha (2009) defines complex predicates as a group of words consisting of a noun (NN/NNP), an adjective (JJ), a verb (VB) or an adverb (RB) followed by a light verb (VBL/VBLI). Light verbs are those which contribute to the tense and agreement of the verb (Butt and Geuder, 2001). However, despite the existence of a light verb tag, it is noticed that in several sentences, verbs followed by auxiliary verbs need to be grouped as a single predicate. Hence, we consider such combinations as belonging to the complex predicate category.

ENG- *According*_VBG *to*_TO *some*_DT *estimates*_NNS *the*_DT *rule*_NN *changes*_NNS *would*_MD ***cut_VB*** *insider*_NN *filings*_NNS *by*_IN *more*_JJR *than*_IN *a*_DT *third*_JJ
URD- [*Kuch*_QN *andaazon*_NN *ke*_CM *mutabiq*_NNCM]_NP [*kanoon*_NN *mein*_CM]_NP [*tabdeeliayan*_NN]_NP[*androni*_JJ *drjbndywn*_NN *ko*_CM]_NP [*ayk*_CD *thayiy*_FR *se*_CM]_NP [***zyada***_I ***kam***_JJ]_JJP [***karey***_VBL ***gi***_AUXT]_VGF

Example 4

Example 4 demonstrates the existence of a light verb in a complex predicate. The English verb *"cut"* is mapped to *"کم کریں گی" (kam karey gi)* belonging to the VBF chunk group.

ENG- *Rolls*_NNP *-_: Royce*_NNP *Motor*_NNP *Cars*_NNPS *Inc.*_NNP ***said***_VBD *it*_PRP ***expects***_VBZ *its*_PRP*$ U.S.*_NNP *sales*_NNS *to*_TO ***remain***_VB *steady*_JJ *at*_IN *about*_IN *1 200*_CD *cars*_NNS *in*_IN *1990*_CD

URD - [*Rolls Royce motor car inc*_NNPC *ne*_CM]_NP [***kaha*_VB**]_VBNF [*wo*_PRP]_NP [*apney*_PRRFP$]_NP [*U.S.*_NNP *ki*_CM]_NP [*frwKt*_NN *ko*_CM]_NP [*1990*_CD *mein*_CM]_NP [*takreeban*_RB]_RBP [*1200_CD karon*_NN *par*_CM]_NP [*mtwazn*_JJ]_JJP [*rakhne*_VBI *ki*_CM]_VGNN [*tawaqqo*_NN]_NP [*karte*_VB *hai*_AUXT]_VGF

Example 5

In example 5, *"said"* corresponds to one Urdu word *"کہا"(kaha)* that also captures the tense information (past). However, consider the verb *"expects"*. This is a clear case of noun-verb complex predicate where *"expects"* is mapped to *"توقع کرتی ہے"(tawaqqo karte hai)*.

ENG- *Not*_RB *all*_PDT *those*_DT *who*_WP ***wrote***_VBD ***oppose***_VBP *the*_DT *changes*_NNS
URD -*wo tamaam jinhon ne* ***likha*** *tabdeeliyon ke* [***mukhalif***_JJ]_JJP [*nahi*_RB]_RBP [***hain***_VBT]_VGF

Example 6

In example 6, verb predicates are *"wrote"* and *"oppose"*. Consider the word *"oppose"*. There are two ways of representing this word in Urdu. As a verb chunk the translation would be *"**mukhalifat** nahi karte"* and as an adjectival chunk *"**mukhalif** nahi hai"*. The latter form of representation is used widely in the available translation corpus. The Urdu equivalent *of "oppose"* is *"مخالف ہیں"(mukhalif hai)*.

Another interesting observation in example 6 is the existence of discontinuous predicates. Though *"oppose"* is one word in English, the Urdu representation has two words that do not occur together. The adverb *"nahi" ~"not"* occurs between the adjective and the verb. Statistically dealing with this issue is extremely challenging and affects the boundaries of other arguments. Generalizing the rules needed to identify discontinuous predicates requires more detailed analysis of the corpus – from the linguistic aspect – and has not been attempted in this paper. We however map *"مخالف نہیں ہیں"(mukhalif nahi hai)* to the predicate *"oppose"*. *"nahi"* is treated as an argument ARG_NEG in PropBank.

4 Projection Results

It is impossible for us to report our projection results on the entire data set as we do not have it manually annotated. For the purpose of evaluation, we manually annotated 100 long sentences (L) and 100 short sentences (S) from the full 2350 sentence set. All the results are reported on

this 200 set of sentences. Set L has sentences that each has more than two verb predicates and several arguments. The number of words per sentence here is greater than 55. S; on the other hand has sentences with about 40 words each and no complex SOV structures.

The results shown in Table 3 are for all tags (verbs+args) that are projected from English onto Urdu. In order to understand why the performance over L dips, consider the results in Table 4 that are for verb projections only. Some long sentences in English have Urdu translations that do not maintain the same structure. For example an English phrase – *"... might **prompt** individuals to **get** out of stocks altogether"* is written in Urdu in a way that the English representation would be *"what makes individuals to **get** out of stocks is ..."*. The Urdu equivalent word for *"prompt"* is missing and the associated lemma gets assigned to the Urdu equivalent of *"get"* (the next lemma). This also affects the argument projections. Another reason is the effect of word alignments itself. Clearly longer sentences have greater alignment errors.

All tags[8]	100 long sentences	100 short sentences
Actual Tags	1267	372
Correct Tags	943	325
Found Tags	1212	353
L : Precision 77.8% Recall 74.4% F-Score 76%		
S: Precision 92% Recall 87.4% F-Score 89.7%		

Table 3: when all tags are considered

Comparing the results of Table 4 to Table 3, we see that argument projections affect the recall. This is because the projections of arguments depend not only on the word alignments but also on the verb predicates. Incorrect verb predicates affect the argument projections.

Only lemma	100 long sentences	100 short sentences
Actual Tags	670	240
Correct Tags	490	208
Overall Tags	720	257
L: Precision 68% Recall 73.1% F-Score 70.45%		
S : Precision 80.9% Recall 86.6% F-Score 83.65%		

Table 4: for verb projections only

Table 5 summarizes the results obtained when only the word alignments are considered to project all tags. But when virtual phrase boundaries in English are also considered, the F-score improves by 8% (Table 6). This is because virtual boundaries in a way mark context switch and when considered in proximity to the word alignments yield better predicate boundaries.

100 long sentences : only alignments	
Actual Tags	1267
Correct Tags	617
Overall Tags	782
Precision 78.9% Recall 48.7% F-Score 60.2%	

Table 5: with only word alignments

100 long sentences : alignments + virtual boundaries	
Actual Tags	1267
Correct Tags	792
Overall Tags	1044
Precision 75.8% Recall 62.5% F-Score 68.5%	

Table 6: with word alignments and virtual boundaries

100 Sentences	ARG 0	ARG 1	ARG 2	ARG 3	ARG M
Long	124	271	67	25	140
Found	111	203	36	12	114
P %	89.5	74.9	53.7	48	81.42
Short	34	47	4	2	19
Found	30	45	4	2	19
P %	88.2	95.7	100	100	100

Table 7: results of argument projections
Precision (P) on arguments

Table 7 shows the results of argument projections over the first 4 arguments of PropBank – ARG0, ARG1, ARG2 and ARG3 (out of 24 arguments, majority are sparse in our test set) and the adjunct tag set ARGM.

5 Automatic Detection

The size of SRL annotated corpus generated for Urdu is limited with only 2350 sentences. To explore the possibilities of augmenting this data set, we train verb predicate and argument detection models. The results show great promise in generating large-scale automatic annotations.

5.1 Verb Predicate Detection

Verb predicate detection happens in two stages. In the first stage, the predicate boundaries are marked using a CRF (Lafferty *et al.*, 2001) based sequence labeling approach. The training data for the model is generated by annotating the automatically annotated Urdu SRL corpus using BI

[8] Tags - lemma (verb predicates) + arguments, Actual tags – number of tags in the English set, Found tags – number of tags transferred to Urdu, Correct Tags – number of tags correctly transferred

annotations. E.g. *kam* B-VG, k*arne par* I-VG. The non-verb predicates are labeled "-1". The model uses POS, chunk and lexical information as features. We report the results on a set of 77 sentences containing a mix of short and long sentences.

Number of verb predicates correctly marked	377
Num of verb predicates found	484
Actual num of verb predicates	451
Precision 77.8% Recall 83.5% F-Score 80.54%	

Table 8: CRF results for verb boundaries

Every verb predicate is associated with a lemma mapped from the English VerbNet map file[9]. E.g. the Urdu verb *"کم کرنے پر" (kam karne par)* has the lemma *"lower"*. The second stage includes assigning these lemmas. Lemma assignment is based on lookups from a VerbNet like map file. We have compiled a large set of Urdu verb predicates by mapping translations found in the automatically annotated corpus to the VerbNet map file. This Urdu verb predicate list also accommodates complex predicates that occur along with verbs such as *"karna – to do", "paana – to get",* etc. (along with different variations of these verbs – *karte, kiya, paate etc.*). This verb predicate list (manually corrected) consists of 800 entries. Since our gold standard test set is very small, the lemma assignment for all verb predicates is 100% (no **pb** values and hence no senses). This list, however, has to be augmented further to meet the standards of the English VerbNet map file.

5.2 Argument Detection

Argument detection (SRL) is done in two steps: (1) argument boundary detection (2) argument label assignment. We perform tests for step 2 to show how well a standard SVM role detection model works on the automatically generated Urdu data set. For each pair of correct predicate *p* and an argument *i* we create a feature representation $F_{p,a}$ ~ set T of all arguments. To train a multi-class role-classifier, given the set T of all arguments, T can be rationalized as $T^{+}_{\arg i}$ (positive instances) and $T^{-}_{\arg i}$ (negative instances) for each argument *i*. In this way, individual ONE-vs-ALL (Gildea and Jurafsky, 2002) classifier for each argument *i* is trained. In the testing phrase, given an unseen sentence, for each argument $F_{p,q}$ is generated and classified by each individual classifier.

We created a set of *standard* SRL features as shown in table 9. The results (Tables 10 and 11), though not impressive, are promising. We believe that by increasing the number of samples (for each argument) in the training set and intelligently controlling the negative samples, the results can be improved significantly.

Training – 2270 sentences with 7315 argument instances. Test – 77 sentences with 496 argument instances. (22 different role types)

BaseLine Features (BL)	phrase-type (syntactic category; NP, PP etc.), predicate (in our case, verb group), path (syntactic path from the argument constituent to the predicate), head words (argument and the predicate respectively), position (whether the phrase is before or after the predicate)
Detailed Features	BL + POS (of the first word in the predicate), chunk tag of the predicate, POS (of the first word of the constituent argument), head word (of the verb group in a complex predicate), named entity (whether the argument contains any named entity, such as location, person, organization etc.)

Table 9: Features for SRL

Kernel/features	Precision	Recall	F-Score
LK – BL	71.88	48.25	57.74
LK – all	73.91	47.55	57.87
PK – BL	74.19	48.25	58.47
PK –all (best)	73.47	49.65	59.26

Table 10: Arg0 performance

Kernel/features	Precision	Recall	F-Score
LK – BL	69.35	22.87	34.40
LK – all	69.84	23.4	35.05
PK – BL	73.77	24.14	36.38
PK –all (best)	73.8	26.06	38.52

Table 11: Arg1 Performances
(PK - polynomial kernel LK – Linear kernel)

6 Conclusion

In this work, we develop an alignment system that is tailor made to fit the SRL problem scope for Urdu. Furthermore, we have shown that despite English being a totally different language, resources for Urdu can be generated if the subtle grammatical nuances of Urdu are accounted for while projecting the annotations. We plan to work on argument boundary detection and explore other features for argument detection. The lemma set generated for Urdu is being refined for finer granularity.

[9] http://verbs.colorado.edu/semlink/semlink1.1/vn-pb/README.TXT

References

Ambati, Vamshi and Chen, Wei, 2007. Cross Lingual Syntax Projection for Resource-Poor Languages. CMU.

Baker, Collin .F., Charles J. Fillmore, John B. Lowe. 1998. The Berkeley Frame Net project. *COLING-ACL*.

Bharati, Akshar, Dipti Misra Sharma, Lakshmi Bai and Rajeev Sangal. 2006. AnnCorra: Annotating Corpora Guidelines For POS And Chunk Annotation For Indian Language. *Technical Report*, Language Technologies Research Centre IIIT, Hyderabad.

Burchardt, Aljoscha and Anette Frank. 2006. Approaching textual entailment with LFG and FrameNet frames. *RTE-2 Workshop*. Venice, Italy.

Butt, Miriam and Wilhelm Geuder. 2001. On the (semi)lexical status of light verbs. *Norbert Corver and Henk van Riemsdijk, (Eds.), Semi-lexical Categories: On the content of function words and the function of content words*, Mouton de Gruyter, pp. 323–370, Berlin.

Diab, Mona and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. *40th Annual Meeting of ACL*, pp. 255-262, Philadelphia, PA.

Dorr, Bonnie, J. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *ACL,* Vol. 20(4), pp. 597-631.

Fillmore, Charles J. 1968. The case for case. *Bach, & Harms(Eds.), Universals in Linguistic Theory*, pp. 1-88. Holt, Rinehart, and Winston, New York.

Fillmore, Charles J. 1982. Frame semantics. *Linguistics in the Morning Calm*, pp.111-137. Hanshin, Seoul, S. Korea.

Fung, Pascale and Benfeng Chen. 2004. BiFrameNet: Bilingual frame semantics resources construction by cross-lingual induction. *20th International Conference on Computational Linguistics*, pp. 931-935, Geneva, Switzerland.

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

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

Hi, Chenhai and Rebecca Hwa. 2005. A backoff model for bootstrapping resources for non-english languages. *Joint Human Language Technology Conference and Conference on EMNLP*, pp. 851-858, Vancouver, BC.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluation translational correspondance using annotation projection. *40th Annual Meeting of ACL*, pp. 392-399, Philadelphia, PA.

Koehn, Phillip. 2005. "Europarl: A parallel corpus for statistical machine translation," MT summit, Citeseer.

Lafferty, John D., Andrew McCallum and C.N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *18th International Conference on Machine Learning*, pp. 282-289.

Liang, Percy, Ben Taskar, and Dan Klein. 2006. Alignment by Agreement, *NAACL*.

Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz. 2004. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, Vol. 19(2), pp. 313-330.

Moschitti, Alessandro. 2008. Kernel methods, syntax and semantics for relational text categorization. *17th ACM CIKM*, pp. 253-262, Napa Valley, CA.

Mukerjee, Amitabh , Ankit Soni and Achala M. Raina. 2006. Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora. *Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pp. 11–18. Sydney.

Mukund, S., Srihari, R. K., and Peterson, E. 2010. An Information Extraction System for Urdu – A Resource Poor Language. *Special Issue on Information Retrieval for Indian Languages*. TALIP.

Pado, Sebastian and Mirella Lapata. 2009. Cross-Lingual annotation Projection of Semantic Roles. *Journal of Artificial Intelligence Research*, Vol. 36, pp. 307-340.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, Vol. 31(1).

Palmer, Martha, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon. 2006. Korean Propbank. *Linguistic data consortium*, Philadelphia.

Philips, Lawrence. 2000. The Double Metaphone Search Algorithm. *C/C++ Users Journal*.

Sinha, R. Mahesh K. 2009. Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus. *ACL International Joint Conference in Natural Language Processing*, pp 40.

Surdeanu, Mihai, Sanda Harabagiu, John Williams and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. *41st Annual Meeting of the Association for Computational Linguistics*, pp. 8-15, Sapporo, Japan.

Taule, Mariona, M. Antonio Marti, and Marta Recasens. 2008. Ancora: Multi level annotated corpora for Catalan and Spanish. *6th International Conference on Language Resources and Evaluation*, Marrakesh, Morocco.

Xue, Nianwen and Martha Palmer. 2009. Adding semantic roles to the Chinese treebank. *Natural Language Engineering*, Vol. 15(1), pp. 143-172.

Yarowsky, David, Grace Ngai and Richard Wicentowski. 2001. Inducing multi lingual text analysis tools via robust projection across aligned corpora. *1st Human Language Technology Conference*, pp. 161-168, San Francisco, CA.