Modeling Selectional Preferences of Verbs and Nouns in String-to-Tree Machine Translation

Maria Nădejde
School of Informatics
University of Edinburgh
m.nadejde@sms.ed.ac.uk

Alexandra Birch
School of Informatics
University of Edinburgh
a.birch@ed.ac.uk

Philipp Koehn
Department of Computer Science
Johns Hopkins University
phi@jhu.edu

Abstract

We address the problem of mistranslated predicate-argument structures in syntax-based machine translation. This paper explores whether knowledge about semantic affinities between the target predicates and their argument fillers is useful for translating ambiguous predicates and arguments. We propose a selectional preference feature based on the selectional association measure of Resnik (1996) and integrate it in a string-to-tree decoder. The feature models selectional preferences of verbs for their core and prepositional arguments as well as selectional preferences of nouns for their prepositional arguments.

We compare our features with a variant of the neural relational dependency language model (RDLM) (Sennrich, 2015) and find that neither of the features improves automatic evaluation metrics. We conclude that mistranslated verbs, errors in the target syntactic trees produced by the decoder and underspecified syntactic relations are negatively impacting these features.

1 Introduction

Syntax-based machine translation systems have had some success when applied to language pairs with major structural differences such as German-English or Chinese-English. Modeling the target side syntactic structure is important in order to produce grammatical, fluent translations and could be an intermediate step on which to build a semantic representation of the target sentence. However these systems still suffer from errors such as scrambled or mis-translated predicate-argument structures. We give a few examples of such errors in Table 1. In example a) the baseline system MT1 mistranslates the verb *besichtigt* as *viewed*. The system MT2 which uses information about the semantic affinity between the verb and its argument produces the correct translation *visited*. The semantic affinity score , shown on the right, for the verb *viewed* and argument *trip* in the syntactic relation *prep_on* is indicating a stronger affinity than for the baseline translation. In example b) the baseline system MT1 mistranslates the noun *Aufnahmen* as *recordings* while the system MT2 produces the correct translation *images* which is a better fit for the prepositional modifier *from the telescope*.

Syntax-based MT systems handle long distance reordering with synchronous translation rules such as:

$$root \rightarrow \langle RB^{\sim 0} VBZ^{\sim 1} sich\; nsubj^{\sim 2} prep^{\sim 3}, \\ RB^{\sim 0} nsubj^{\sim 2} VBZ^{\sim 1} prep^{\sim 3} \rangle$$

This rule is useful for reordering the verb and its arguments according to the target side word order. However the rule does not contain the lexical head for the verb, the subject and the prepositional modifier. Therefore the entire predicate argument structure is translated by subsequent independent rules. The language model context will capture at most the verb and one main argument. Due to the lack of a larger source or target context the resulting predicate-argument structures are often not semantically coherent.

This paper explores whether knowledge about semantic affinities between the target predicates and their argument fillers is useful for translating ambiguous predicates and arguments. We propose a selectional preference feature for string-to-tree statistical machine translation based on the information theoretic measure of Resnik (1996). The feature models selectional preferences of verbs for

Proceedings of the First Conference on Machine Translation, Volume 1: Research Papers, pages 32–42,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

			(relation, predicate, argument)	Affinity
a)	SRC	Bei nur einer Reise können nicht alle davon *besichtigt* werden.		
	REF	You won't be able to *visit* all of them on one trip .		
	MT1	Not all of them can be *viewed* on only one trip.	(prep_on, *viewed*, trip)	-0.154
	MT2	Not all of them can be *visited* on only one trip.	(prep_on, *visited*, trip)	1.042
b)	SRC	Eine der schärfsten *Aufnahmen* des Hubble-Teleskops		
	REF	One of the sharpest *pictures* from the Hubble telescope		
	MT1	One of the strongest *recordings* of the Hubble telescope	(prep_of, *recordings*, telescope)	-0.0004
	MT2	One of the strongest *images* from the Hubble telescope	(prep_from, *images*, telescope)	0.3917

Table 1: Examples of errors in the predicate-argument structure produced by a syntax-based MT system. a) mistranslated verb b) mistranslated noun. Semantic affinity scores are shown on the right. Higher scores indicate a stronger affinity. Negative scores indicate a lack of affinity.

their core and prepositional arguments as well as selectional preferences of nouns for their prepositional arguments.

Previous work has addressed the selectional preferences of prepositions for noun classes (Weller et al., 2014) but not the semantic affinities between a predicate and its argument class. Another line of research on improving translation of predicate-argument structures includes modeling reordering and deletion of semantic roles (Wu and Fung, 2009; Liu and Gildea, 2010; Li et al., 2013). These models however do not encode information about the lexical semantic affinities between target predicates and their arguments. Sennrich (2015) proposes a relational dependency language model (RDLM) for string-to-tree machine translation. One component of RDLM predicts the head word of a dependent conditioned on a wide syntactic context. Our feature is different as it quantifies the amount of information that the predicate carries about the argument class filling a particular syntactic function.

For one variant of the proposed feature we found a slight improvement in automatic evaluation metrics when translating short sentences as well as an increase in precision for verb translation. However the features generally did not improve automatic evaluation metrics. We conclude that mistranslated verbs, errors in the target syntactic trees produced by the decoder and underspecified syntactic relations are negatively impacting these features.

The paper is structured as follows. Section 2 describes related work on improving translation of predicate-argument structures. Section 3 introduces the selectional preference feature. Section 4 describes the experimental setup and Section 5 presents the results of automatic evaluation as well as a qualitative analysis of the machine translated output.

2 Related work

From a syntactic perspective, a correct predicate-argument structure will have the sub-categorization frame of the predicate filled in. Weller et al. (2013) use sub-categorization information to improve case-prediction for noun phrases when translating into German. Case prediction for noun phrases is important in the German language as it indicates the grammatical function. Their approach however did not produce strong improvements over the baseline. From a large corpus annotated with dependency relations, they extract verb-noun tuples and their associated syntactic functions: direct object, indirect object, subject. They also extract triples of verb-preposition-noun in order to predict the case of noun-phrases within prepositional-phrases. The probabilities of such tuples and triples are computed using relative frequencies and then used as a feature for a CRF classifier that predicts the case of noun-phrases. Weller et al. (2013) apply the CRF classifier to the output of a word-to-stem phrased-based translation system as a post-processing step. In contrast, our model is used directly as a feature in the decoder. While Weller et al. (2013) identify the arguments of the verb and their grammatical function by projecting the information from the source sentence we use the dependency tree produced by the string-to-tree decoder. We also consider prepositional modifiers of nouns.

Weller et al. (2014) propose using noun class information to model selectional preferences of prepositions in a string-to-tree translation system. They use the noun class information to annotate PP translation rules in order to restrict their applicability to specific semantic classes. In our work we don't impose hard constraints on the translation rules, but rather soft constraints using our model as a feature in the decoder. While we use word embeddings to cluster arguments, Weller et al. (2014) experiment with a lexical semantic taxonomy and clustering words based on cooccurrences within a window or syntactic features extracted from dependency-parsed data.

Modeling reordering and deletion of semantic roles (Wu and Fung, 2009; Liu and Gildea, 2010; Li et al., 2013) has been another line of research on improving translation of predicate-argument structures. Liu and Gildea (2010) propose modeling reordering of a complete semantic frame while Li et al. (2013) propose finer grained features that distinguish between predicate-argument reordering and argument-argument reordering. Gao and Vogel (2011) and Bazrafshan and Gildea (2013) annotate target non-terminals with the semantic roles they cover in order to extract synchronous grammar rules that cover the entire predicate argument structure. These models however do not encode information about the lexical semantic affinities between target predicates and their arguments.

In this work we focus on using selectional preference over predicate and arguments in the target as this is a simple way of leveraging external knowledge in the translation framework.

3 Selectional Preference Feature

3.1 Learning Selectional Preferences

Selectional preferences describe the semantic affinities between predicates and their argument fillers. For example, the verb "drinks" has a strong preference for arguments in the conceptual class of "liquids". Therefore the word "wine" can be disambiguated when it appears in relation to the verb "drinks". A corpus driven approach to modeling selectional preferences usually involves extracting triples of (*syntactic relation, predicate, argument*) and computing co-occurrence statistics. The predicate and argument are represented by their head words and the triples are extracted from automatically parsed data. Another typical step is generalizing over seen arguments. Approaches to generalization include using an ontology such as WordNet (Resnik, 1996), using distributional semantics similarity (Erk et al., 2010; Séaghdha, 2010; Ritter et al., 2010), clustering (Sun and Korhonen, 2009), multi-modal datasets (Shutova et al., 2015), and neural networks (Cruys, 2014).

Our feature is based on the measure proposed by Resnik (1996). It uses unsupervised clusters to generalize over seen arguments. Resnik (1996) uses selectional preferences of predicates for word sense disambiguation. The information theoretic measure for selectional preference proposed by Resnik quantifies the difference between the posterior distribution of an argument class given the verb and the prior distribution of the class. For instance, "person" has a higher prior probability than "insect" to appear in the subject relation, but, knowing the verb is "fly", the posterior probability becomes higher for "insect".

Resnik's model defines *selectional preference strength* of a predicate as:

$$\begin{aligned} SelPref(p,r) &= KL(P(c|p,r) \parallel P(c|r)) \\ &= \sum_{c} P(c|p,r) log \frac{P(c|p,r)}{P(c|r)} \end{aligned} \quad (1)$$

where KL is the Kullback - Leibler divergence, r is the relation type, p is the predicate and c is the conceptual class of the argument. Resnik uses WordNet to obtain the conceptual classes of arguments, therefore generalizing over seen arguments. The *selectional association* or semantic affinity between a predicate and an argument class is quantified as the relative contribution of the class towards the overall selectional strength of the predicate:

$$SelAssoc(p,r,c) = \frac{P(c|p,r) log \frac{P(c|p,r)}{P(c|r)}}{SelStr(p,r)} \quad (2)$$

We give examples of the *selectional preference strength* and *selectional association* scores for different verbs and their arguments in Table 2. The verb *see* takes on many arguments as direct objects and therefore has a lower selectional preference strength for this syntactic relation. In contrast the predicate *hereditary* takes on fewer arguments for which it has a stronger selectional preference.

Several selectional preference models have been used as features in discriminative syntactic parsing systems. Cohen et al. (2012) observe

Verb	Relation	SelPref	Argument	SelAssoc
see	dobj	0.56	PRN	0.123
			movie	0.022
			episode	0.001
is–hereditary	nsubj	1.69	disease	0.267
			monarchy	0.148
			title	0.082
drink	dobj	3.90	water	0.144
			wine	0.061
			glass	0.027

Table 2: Example of *selectional preference* (SelPref) and *selectional association* (SelAssoc) scores for different verbs. PRN is the class of pronouns.

that when parsing out-of-domain data many attachment errors occur for the following syntactic configurations: head (V or N) – prep – obj and head (N) – adj. The authors proposed a class-based measure of selectional preferences for these syntactic configurations and learn the argument classes using Latent Dirichlet Allocation (LDA). Kiperwasser and Goldberg (2015) compare different measures of lexical association between head word and modifier word for improving dependency parsing. Their results show that the association measure based on pointwise mutual information (PMI) has similar generalization capabilities as a measure of distributional similarity between word embeddings. van Noord (2007) has shown that bilexical association scores computed using PMI for all types of dependency relations are a useful feature for improving dependency parsing in Dutch.

3.2 Adaptation of Selectional Preference Models for Syntax-Based Machine Translation.

We are interested in modeling selectional preferences of verbs for their core and prepositional arguments as well as selectional preferences of nouns for their prepositional arguments. We identify the relation between a predicate and its modifier from the dependency tree produced by a string-to-tree machine translation system. Since we are interested in using the feature during decoding, we need the model to be fast to query and have broad coverage.

Our selectional preference feature is a variant of the information theoretic measure of Resnik (1996) defined in Eq 2. While Resnik uses the WordNet classes of the arguments, this is not appropriate for a machine translation task where the vocabulary has millions of words and English is not the only targeted language. Therefore we adapt Resnik's selectional association measure in two ways.

In the first model *SelAssoc_L* we compute the co-occurrence statistics defined in Eq 2 over lemmas of the predicate and argument head words.

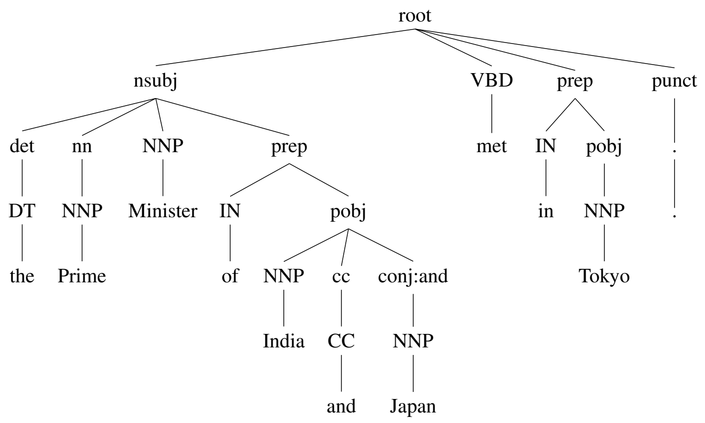
In the second model *SelAssoc_C* we replace the WordNet classes in Eq 2 with word clusters[1] . We obtain the word clusters by applying the k-means algorithm to the glovec word embeddings (Pennington et al., 2014).

Prepositional phrase attachment remains a frequent and challenging error for syntactic parsers (Kummerfeld et al., 2012) and translation of prepositions is a challenge for SMT (Weller et al., 2014). Therefore we decide to use two separate features: one for main arguments (*nsubj, nsubjpass, dobj, iobj*) and one for prepositional arguments.

3.3 Comparison with a Neural Relational Dependency Language Model.

Sennrich (2015) proposes a relational dependency language model (RDLM) for string-to-tree machine translation, which he trains using a feed-forward neural network. For a sentence S with symbols $w_1, w_2, ...w_n$ and dependency labels $l_1, l_2, ...l_n$ with l_i the label of the incoming arc at position i, RDLM is defined as:

[1] We have not done experiments with WordNet classes.



relation	predicate	argument
nsubj	met	Minister
prep_in	met	Tokyo
prep_of	Minister	India

Figure 1: Example of a translation and its dependency tree in constituency representation produced by the string-to-tree SMT system. Triples extracted during decoding are shown on the right.

$$
\begin{aligned}
&P(S, D) \approx \prod_{i=1}^{n} P_l(i) \times P_w(i) \\
&P_l(i) = P(l_i \mid h_s(i)_1^q, l_s(i)_1^q, h_a(i)_1^r, l_a(i)_1^r) \\
&P_w(i) = P(w_i \mid h_s(i)_1^q, l_s(i)_1^q, h_a(i)_1^r, l_a(i)_1^r, l_i)
\end{aligned}
\tag{3}
$$

where for each of q siblings and r ancestors of w_i, h_s and h_a are their head words and l_s and l_a their dependency labels. The $P_w(i)$ distribution models similar information as our proposed feature $SelAssoc$. However we use $h_a(i)_1, l_i$ as context and consider only a subset of dependency labels: *nsubj, nsubjpass, dobj, iobj, prep*. The reduced context alleviates problems of data sparsity and is more reliably extracted at decoding time. The subset of dependency relations identify arguments for which predicates might exhibit selectional preferences. Our feature is different from $RDLM - P_w$ as it quantifies the amount of information that the predicate carries about the argument class filling a particular syntactic function. We hypothesize that such information is useful when translating arguments that appear less frequently in the training data but are prototypical for certain predicates. For example the triples *(bus, drive, dobj)* and *(van, drive, dobj)* have the following log posterior probabilities and $SelAssoc$ scores: log P(bus | drive, dobj) = -5.44, log P(van | drive, dobj)= -5.58 and SelAssoc(bus, drive, dobj) = 0.0079, SelAssoc(van, drive, dobj) = 0.0103.

4 Experimental setup

Our baseline system for translating German into English is the Moses string-to-tree toolkit implementing GHKM rule extraction (Galley et al., 2004, 2006; Williams and Koehn, 2012). The string-to-tree translation model is based on a synchronous context-free grammar (SCFG) that is extracted from word-aligned parallel data with target-side syntactic annotation. The system was trained on all available data provided at WMT15 [2] (Bojar et al., 2015). The number of sentences in the training, tuning and test sets are shown in Table 3. We use the following rule extraction parameters: *Rule Depth = 5, Node Count = 20, Rule Size = 5*. At decoding time we give a high penalty to glue rules and allow non-terminals to span a maximum of 50 words. We train a 5-gram language model on all available monolingual data [3] using the SRILM toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing (Chen and Goodman, 1998) for training and KenLM (Heafield, 2011) for language model scoring during decoding.

Train	Tune	Test
4,472,694	2000	8172

Table 3: Number of sentences in the training, tuning and test sets. The test set consists of the WMT newstest2013, 2014 and 2015.

The English side of the parallel corpus is annotated with dependency relations using the Stanford dependency parser (Chen and Manning, 2014). The dependency structure is then converted to a constituency representation which is needed to run the GHKM rule extraction. We use the conversion

[2] http://www.statmt.org/wmt15/translation-task.html

[3] target side of the parallel corpus, the monolingual English News Crawl, Gigaword and news-commentary

algorithm and the head word extraction method described in Sennrich (2015).

For training the selectional preference features we extract triples of *(dependency relation, predicate, argument)* from parsed data, where the predicate and argument are identified by their head word. We use the english side of the parallel data and the Gigaword v.5 corpus parsed with Stanford typed dependencies (Napoles et al., 2012). We use Stanford dependencies in the collapsed version which resolves coordination [4] and collapses the prepositions. Figure 1 shows an example of a translated sentence, its dependency tree produced by the string-to-tree system and the triples extracted at decoding time. We consider the following main arguments: *nsubj, nsubjpass, dobj, iobj* and *prep* arguments attached to both verbs and nouns. Table 4 shows the number of extracted triples.

Type of relation	Number of triples
main	540,109,283
prep	810,118,653
nsubj	315,852,775
nsubjpass	32,111,962
dobj	188,412,178
iobj	3,732,368

Table 4: Number of relation triples extracted from parsed data. The data consists of the English side of the parallel data and Gigaword. *main* arguments include: nsubj, nsubjpass, dobj, iobj.

We integrate the feature in a bottom-up chart decoder. The feature has several scores:

- A counter for the dependency triples covered by the current hypothesis.
- A selectional association score aggregated over all main arguments: nsubj, nsubjpass, dobj, iobj.
- A selectional association score aggregated over all prepositional arguments with no distinction between noun and verb modifiers.

For both tuning and evaluation of all machine translation systems we use a combination of the cased BLEU score and head-word chain metric (HWCM) (Liu and Gildea, 2005). The HWCM metric implemented in the Moses toolkit computes the harmonic mean of precision and recall over head-word chains of length 1 to 4. The head-word chains are extracted directly from the dependency tree produced by the string-to-tree decoder and from the parsed reference. Tuning is performed using batch MIRA (Cherry and Foster, 2012) on 1000-best lists. We report evaluation scores averaged over the newstest2013, newstest2014 and newstest2015 data sets provided by WMT15.

5 Evaluation

5.1 Error analysis

We wanted to get an idea about how often the verb and its arguments are mistranslated. For this purpose we manually annotated errors in sentences with more than 5 words and at most 15 words. With this criterion we avoided translations with scrambled predicate-argument structures. Each sentence had roughly one main verb.

To have a more reliable error annotation we first post-edited 100 translations from the baseline system. We then compared the translations with their post-editions and annotated error categories using the BLAST tool (Stymne, 2011). We considered a *sense* error category when there was a wrong lexical choice for the head of a main argument, a prepositional modifier or the main verb. We also annotated mistranslated prepositions.

Error Category	Error Count	Total
Preposition	18	143
Sense	53	388
Main argument	18	145
Prep modifier	9	143
Main verb	26	100

Table 5: Number of mistranslated words in 100 sentences manually annotated with error categories.

In Table 5 we can see that 26 percent of the verbs are mistranslated and about 10 percent of the arguments. Mistranslated verbs are problematic since the feature produces the selectional association scores for the wrong verb. Although the semantic affinity is mutual, the formulation of the score conditions on the verb. In the cases when both the verb and the argument are mistranslated the association score might be high although the translation is not faithful to the source.

[4] Coordination is not resolved at decoding time.

5.2 Evaluation of the Selectional Preference Feature

First, we determine the effectiveness of our selectional association features. We compare the two different selectional association features described in section 3.2: *SelAssoc_L* and *SelAssoc_C* . We report the results of automatic evaluation in Table 6.

Neither of the features improved the automatic evaluation scores. The *SelAssoc_L* suffers from data sparsity while the *SelAssoc_C* feature is overgeneralizing due to noisy clustering. Adding both features compensates for these issues, however we only see a slight improvement in BLEU scores for shorter sentences[5]: 25.59 compared to 25.40 for the baseline system. We further investigate whether sparse features are more informative.

System	BLEU -c	HWCM
Baseline	26.45	24.47
+ SelAssoc_L	$26.41_{-.04}$	$24.52_{+.05}$
+ SelAssoc_C	$26.48_{+.03}$	$24.54_{+.07}$
+ SelAssoc_L + SelAssoc_C	$26.48_{+.03}$	$24.47_{+.00}$
+ Bin (SelAssoc_L + SelAssoc_C)	$26.37_{-.08}$	$24.53_{+.06}$
+ RDLM–P_w (1, 0, 0)	$26.35_{-.10}$	$24.75_{+.28}$
+ RDLM–P_w (2, 1, 1)	$26.38_{-.07}$	$24.83_{+.36}$

Table 6: Results for string-to-tree systems with *SelAssoc* and RDLM–P_w features. The number of clusters used with *SelAssoc_C* is 500. The triples in parenthesis indicate the context size for ancestors, left siblings and right siblings respectively. The RDLM–P_w configuration (1, 0, 0) captures similar syntactic context as the selectional preference features.

We changed the format of the features in order to experiment with sparse features. By using sparse features we let the tuning algorithm discriminate between low and high values of the *SelAssoc* score. For each of the *SelAssoc* features we normalized the scores to have zero mean and standard deviation one and mapped them to their corresponding percentile. A sparse feature was created for each percentile, below and above the mean [6] resulting in a total of 20 sparse features. However this formulation of the feature also did not improve the evaluation scores as shown in the fifth row of Table 6.

The lack of variance in automatic evaluation scores can be explained by: a) the feature touches only a few words in the translation and b) the relation between a predicate and its argument is identified at later stages of the bottom-up chart-based decoding when many lexical choices have already been pruned out. The *SelAssoc* scores, similar to mutual information scores, are sensitive to outlier events with low frequencies in the training data. In the next section we investigate whether a more robust model would mitigate some of these issues and experiment with a neural relational dependency language model (RDLM) (Sennrich, 2015).

5.3 Comparison with a Relational Dependency LM

The RDLM (Sennrich, 2015) is a feed-forward neural network which learns two probability distributions conditioned on a large syntactic context described in Eq 3: P_w predicts the head word of the dependent and P_l the dependency relation. We compare our feature with RDLM–P_w.

For training the RDLM–P_w we use the parameters for the feed-forward neural network described in Sennrich (2015): 150 dimensions for input layer, 750 dimensions for the hidden layer, a vocabulary of 500 000 words and 100 noise samples. We train the RDLM–P_w on the target side of the parallel data. Although we use less data than for training the *SelAssoc* features, the neural network is inherently good at learning generalizations and selecting the appropriate conditioning context.

We experiment with different configurations for RDLM–P_w by varying the number of ancestors as well as left and right siblings:

- ancestors = 1, left = 0, right = 0
- ancestors = 2, left = 1, right = 1

The first configuration captures similar syntactic context as the *SelAssoc* features. The only exception is the *prep* relation for which the head of *pobj*, the actual preposition, is a sibling of the argument. The results are shown in the last two lines of Table 6 and the configuration is marked between parentheses for the ancestors, left siblings and right siblings respectively.

The RDLM–P_w performs slightly better than the selectional preference feature in terms of the HWCM scores. An increase in HWCM is to be

[5] 2701 sentences with more than 5 words and at most 15 words

[6] Up to two standard deviations below the mean and three standard deviations above the mean.

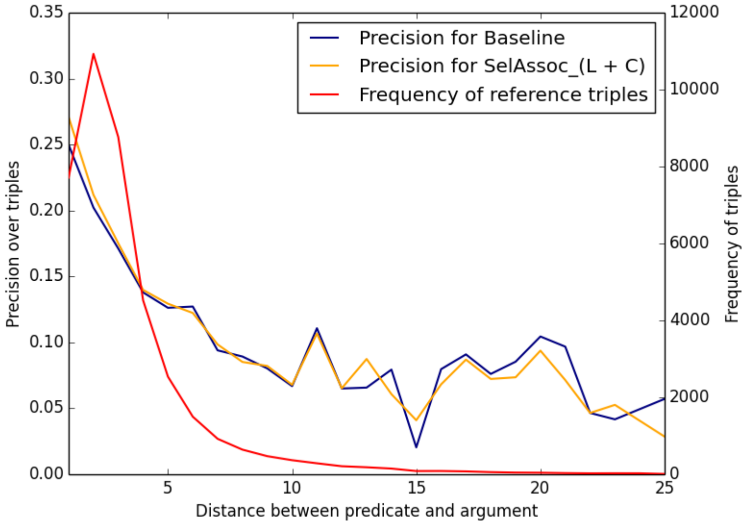


Figure 2: Frequency and translation precision of triples with respect to the distance between the predicate and its arguments. Frequency is computed for triples extracted from the reference sentences of the tests sets. Translation precision is computed over triples extracted from the output of the two translation systems: baseline system and the system with *SelAssoc_L* and *SelAssoc_C* features.

expected since the RDLM–P_w models all dependency relations. However there is not a significant contribution from having a larger syntactic context.

5.4 Analysis

In this section we investigate possible reasons for the low impact of our selectional preference features. We look at how frequently our features are triggered, and how precision is influenced by the distance between predicates and arguments.

Firstly we are interested in how often the feature triggers and how it influences the overall selectional association score of the test set. On average, 4.85 triples can be extracted per sentence produced by our system. Out of these, 4.35 triples get scored by the *SelAssoc_C* feature and 3.56 by the *SelAssoc_L* feature. The selectional association scores are higher on average for our system than for the baseline as shown in Table 7. The *SelAssoc_C* feature seems to overgeneralize for the *prep* relations as the scores are on average higher than for the reference triples. We therefore conclude that our feature is having an impact on the translation system.

Secondly we want to understand the interaction between the *SelAssoc* features and the language model. For this purpose we compute the frequency and translation precision of triples with respect to the distance between the predicate and its arguments. Figure 2 shows the frequency of triples extracted from the reference sentence as well as the translation precision of triples extracted from the output of the translation systems. For more reliable precision scores we lemmatized all predicates and arguments. Most arguments are within a 5 word window from the predicate. Therefore most triples are also scored by the language model. For these triples we see only a slight increase in precision for our system. This result indicates that for predicates and arguments that are close to each other, the feature is not adding much information. As the distance increases the precision decreases drastically for both systems. A longer distance between predicates and arguments also implies a

	SelAssoc_L		SelAssoc_C	
System	main	prep	main	prep
Baseline	0.067	0.039	0.164	0.147
+ SelAssoc_L + SelAssoc_C	0.074	0.041	0.175	0.305
Reference	0.077	0.043	0.186	0.163

Table 7: Average selectional association scores for the test sets. Scores are aggregated over the *main* and *prep* argument types. *main* arguments include: nsubj, nsubjpass, dobj, iobj.

Source	Das 16-jährige Mädchen und der 19-jährige Mann **brachen** kurz nach Sonntagmittag in Govetts Leap in Blackheath **zu ihrer Tour auf**.
Reference	The 16-year old girl and the 19-year old man **went on their tour** shortly after Sunday lunch at Govetts Leap in Blackheath.
Baseline	The 16-year old girl and the 19-year old man **broke** shortly after Sunday lunch in Govetts Leap in Blackheath **on their tour**.

Figure 3: Examples of a complex sentence with multiple prepositional modifiers. Information about semantic roles is needed to identify the relevant prepositional modifier.

more complex syntactic structure which will negatively impact the quality of extracted triples and the selectional association scores.

5.5 Discussion

One reason for the small impact of both *SelAssoc* and RDLM–P_w features could be the poor quality of the syntactic trees produced by the decoder for longer sentences. In the cases where the relation between predicate and argument can be reliably extracted, such as the example in Fig 1, the features are not adding more information than is already covered by the language model.

In more complex sentences there are cases where the features score modifiers that are not important for disambiguating the verb. The example in Figure 3 has several prepositional modifiers but only *on tour* could help disambiguate the verb *brachen (went)*. In such cases identifying the semantic roles of the modifiers in the source and projecting them on the target might be useful for better estimation of semantic affinities.

The error analysis on short sentences showed that translation of verbs is problematic for syntax-based systems. This is confirmed by the low precision scores[7] for verb translation shown in Table 8. Although there is a slight improvement in precision, generally mistranslated verbs impact our features as the semantic affinity is scored for the wrong verb. A solution would be to add the source verbs in the conditioning context.

System	Precision
baseline	46.10
+ SelAssoc_L + SelAssoc_C	$46.26_{+.16}$
+ RDLM–P_w (2, 1, 1)	$46.31_{+.21}$

Table 8: Evaluation of verb translation in the test set. Precision scores are computed over verb lemmas against the reference translations.

6 Conclusions

This paper explores whether knowledge about semantic affinities between the target predicates and their argument fillers is useful for translating ambiguous predicates and arguments. We propose three variants of a selectional preference feature for string-to-tree statistical machine translation based on the selectional association measure of Resnik (1996). We compare our features with a variant of the neural relational dependency language model (RDLM) (Sennrich, 2015) and find that neither of the features improves automatic evaluation metrics. We conclude that mistranslated verbs, errors in the target syntactic trees produced by the decoder and underspecified syntactic relations are negatively impacting these features. We propose to address these issues in future work by augmenting the feature with source side information such as the source verb and the semantic roles of its arguments.

Acknowledgments

We thank the anonymous reviewers as well as Rico Sennrich for his feedback and assistance with RDLM . This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 644402 (HimL) and 645452 (QT21). We are also grateful for support by a Google Faculty Research Award.

References

Marzieh Bazrafshan and Daniel Gildea. 2013. Semantic roles for string to tree machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, ACL 2013, pages 419–423.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi.

[7] The precision scores were computed over verb lemmas extracted automatically from the test sets. In total 21633 source verbs were evaluated.

2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pages 1–46.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 740–750.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report, Harvard University.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montreal, Canada, NAACL HLT '12, pages 427–436.

Raphael Cohen, Yoav Goldberg, and Michael Elhadad. 2012. Domain adaptation of a dependency parser with a class-class selectional preference model. In *Proceedings of ACL 2012 Student Research Workshop*. ACL '12, pages 43–48.

Tim Van De Cruys. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL '14, pages 26–35.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Comput. Linguist.* 36(4):723–763.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Morristown, NJ, USA, pages 961–968.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*. HLT-NAACL '04.

Qin Gao and Stephan Vogel. 2011. Utilizing target-side semantic role labels to assist hierarchical phrase-based machine translation. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Portland, Oregon, USA, pages 107–115.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland, WMT '11, pages 187–197.

Eliyahu Kiperwasser and Yoav Goldberg. 2015. Semi-supervised dependency parsing using bilexical contextual features from auto-parsed data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 1348–1353.

Jonathan K. Kummerfeld, David Hall, James R. Curran, and Dan Klein. 2012. Parser showdown at the wall street corral: An empirical investigation of error types in parser output. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL '12, pages 1048–1059.

Junhui Li, Philip Resnik, and Hal Daum. 2013. Modeling syntactic and semantic structures in hierarchical phrase-based translation. In *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics,*. Atlanta, Georgia, USA, number June in NAACL-HLT 2013, pages 540–549.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, MI, pages 25–32.

Ding Liu and Daniel Gildea. 2010. Semantic role features for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*. pages 716–724.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for

Computational Linguistics, Stroudsburg, PA, USA, AKBC-WEKEX '12, pages 95–100.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. pages 1532–1543.

Philip Resnik. 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition* 61:127–159.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA, ACL '10, pages 424–434.

Diarmuid Ó. Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA, ACL '10, pages 435–444.

Rico Sennrich. 2015. Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation. *Transactions of the Association for Computational Linguistics* 3:169–182.

Ekaterina Shutova, Niket Tandon, and Gerard de Melo. 2015. Perceptually grounded selectional preferences. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. ACL '15, pages 950–960.

Andreas Stolcke. 2002. SRILM – an Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*. volume 3.

Sara Stymne. 2011. Blast: A tool for error analysis of machine translation output. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*. Portland, Oregon, HLT '11, pages 56–61.

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*. EMNLP '09, pages 638–647.

Gertjan van Noord. 2007. Using self-trained bilexical preferences to improve disambiguation accuracy. In *Proceedings of the 10th International Conference on Parsing Technologies*. IWPT '07, pages 1–10.

Marion Weller, Alexander Fraser, and Sabine Schulte im Walde. 2013. Using subcategorization knowledge to improve case prediction for translation to german. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 593–603.

Marion Weller, Sabine Schulte Im Walde, and Alexander Fraser. 2014. Using noun class information to model selectional preferences for translating prepositions in smt. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*. Vancouver, BC, AMTA '14, pages 275–287.

Philip Williams and Philipp Koehn. 2012. Ghkm rule extraction and scope-3 parsing in moses. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. pages 388–394.

Dekai Wu and Pascale Fung. 2009. Semantic roles for SMT: a hybrid two-pass model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. pages 13–16.