Shallow Semantic Trees for SMT

Wilker Aziz, Miguel Rios and **Lucia Specia**
Research Group in Computational Linguistics
University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB, UK
{w.aziz, m.rios, l.specia}@wlv.ac.uk

Abstract

We present a translation model enriched with shallow syntactic and semantic information about the source language. Base-phrase labels and semantic role labels are incorporated into an hierarchical model by creating shallow semantic "trees". Results show an increase in performance of up to 6% in BLEU scores for English-Spanish translation over a standard phrase-based SMT baseline.

1 Introduction

The use of semantic information to improve Statistical Machine Translation (SMT) is a very recent research topic that has been attracting significant attention. In this paper we describe our participation in the shared translation task of the 6th Workshop on Statistical Machine Translation (WMT) with a system that incorporates shallow syntactic and semantic information into hierarchical SMT models.

The system is based on the Moses toolkit (Hoang et al., 2009; Koehn et al., 2007) using hierarchical models informed with shallow syntactic (chunks) and semantic (semantic role labels) information for the source language. The toolkit SENNA (Collobert et al., 2011) is used to provide base-phrases (chunks) and semantic role labels.

Experiments with English-Spanish and English-German news datasets show promising results and highlight important issues about the use of semantic information in hierarchical models as well as a number of possible directions for further research.

The remaining of the paper is organized as follows: Section 2 presents related work; Section 3 describes the method; Section 4 presents the results obtained for the English-Spanish and English-German translation tasks; and Section 5 brings some conclusions and directions for further research.

2 Related Work

In hierarchical SMT (Chiang, 2005), a Synchronous Context Free Grammar (SCFG) is learned from a parallel corpus.The model capitalizes on the recursive nature of language replacing sub-phrases by an unlabeled nonterminal. Hierarchical models are known to produce high coverage rules, once they are only constrained by the word alignment. Nevertheless the lack of specialized vocabulary also leads to spurious ambiguity (Chiang, 2005).

Syntax-based models are hierarchical models whose rules are constrained by syntactic information.The syntactic constraints have an impact in the rule extraction process, reducing drastically the number of rules available to the system. While this may be helpful to reduce ambiguity, it can lead to poorer performance (Ambati and Lavie, 2008).

Motivated by the fact that syntactically constraining a hierarchical model can decrease translation quality, some attempts to overcome the problems at rule extraction time have been made. Venugopal and Zollmann (2006) propose a heuristic method to relax parse trees known as Syntax Augmented Machine Translation (SAMT). Significant gains are obtained by grouping nonterminals under categories when they do not span across syntactic constituents.

Hoang and Koehn (2010) propose a soft syntax-based model which combines the precision of a syntax-constrained model with the coverage of an

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 316–322,
Edinburgh, Scotland, UK, July 30–31, 2011.

unconstrained hierarchical model. Instead of having heuristic strategies to combine nonterminals in a parse tree, whenever a rule cannot be retrieved because it does not span a constituent, the extraction procedure falls back to the hierarchical approach, retrieving a rule with unlabeled nonterminals. Performance gains are reported over standard hierarchical models using both full parse trees and shallow syntax.

Moving beyond syntactic information, some attempts have recently been made to add semantic annotations to SMT. Wu and Fung (2009) present a two-pass model to incorporate semantic information to the phrase-based SMT pipeline. The method performs conventional translation in a first step, followed by a constituent reordering step seeking to maximize the cross-lingual match of the semantic role labels of the translation and source sentences.

Liu and Gildea (2010) add features extracted from the source sentences annotated with semantic role labels in a tree-to-string SMT model. They modify a syntax-based SMT system in order to penalize/reward role reordering and role deletion. The input sentence is parsed for semantic roles and the roles are then projected onto the target side using word alignment information at decoding time. They assume that a one-to-one mapping between source and target roles is desirable.

Baker et al. (2010) propose to graft semantic information, namely named entities and modalities, to syntactic tags in a syntax-based model. The vocabulary of nonterminals is specialized using the semantic categories, for instance, a noun phrase (NP) whose head is a geopolitical entity (GPE) will be tagged as NPGPE, making the rule table less ambiguous.

Similar to (Baker et al., 2010) we specialize a vocabulary of syntactic nonterminals with semantic information, however we use shallow syntax (base-phrases) and semantic role labels instead of constituent parse and named entities. The resulting shallow trees are relaxed following SAMT (Venugopal and Zollmann, 2006). Different from previous work we add the semantic knowledge at the level of the corpus annotation. As a consequence, instead of biasing deletion and reordering through additional features (Liu and Gildea, 2010), we learn hierarchical rules that encode those phenomena, taking also into account the semantic role of base-phrases.

3 Proposed Method

The proposed method is based on an extension of the hierarchical models in Moses using source language information. Our submission included systems for two language pairs: English-Spanish (en-es) and English-German (en-de) and was constrained to using data provided by WMT11. Phrase and rule extraction were performed using the entire en-es and en-de portions of Europarl. Model parameters were tuned using the *news-test2008* dataset. Three 5-gram Spanish and German language models were trained using SRILM[1] with the News Commentaries ($\sim$ 160K sentences), Europarl ($\sim$ 2M sentences) and News ($\sim$ 5M sentences) corpora. These models were interpolated using scripts provided in Moses (Koehn and Schroeder, 2007).

At pre-processing stage, sentences longer than 80 tokens were filtered from the training/development corpus. The parallel corpus was then tokenized and truecased. Additionally, for en-de, compound splitting of the German side of the corpus was performed using a frequency based method described in (Koehn and Knight, 2003). This method helps alleviate sparsity, reducing the size of the vocabulary by decomposing compounds into their base words. Recasing and detokenization, along with compound merging of the translations into German, were handled at post-processing stage. Compound merging was performed by finding the most likely sequences of words to be merged into previously seen compounds (Stymne, 2009).

3.1 Source Language Annotation

For rule extraction, training and test, the English side of the corpus was annotated with Semantic Role Labels (SRL) using the toolkit SENNA[2], which also outputs POS and base-phrase (without prepositional attachment) tags. The resulting source language annotation was used to produce trees in order to build a tree-to-string model in Moses.

[1] http://www.speech.sri.com/projects/srilm/

[2] http://ml.nec-labs.com/senna/

S															
NP	VP			NP		PP	NP	O	O	NP	VP			NP	ADVP
PRP	VBZ	TO	VB	DT	NN	TO	NN	PUNC	CC	PRP	VBZ	RB	VBD	WDT	RB
he	intends	to	donate	this	money	to	charity	,	but	he	has	not	decided	which	yet

Figure 1: Example of POS tags and base-phrase annotation. Base-phrases: noun-phrase (NP), verb-phrase (VP), prepositional-phrase (PP), adverbial-phrase (ADVP), outside-of-a-phrase (O)

In order to derive trees for the source side of the corpus from this annotation, a new level is created to add the POS tags for each word form. Syntactic tags are then added by grouping words and POS tags into base phrases using linguistic information as given by SENNA. Figure 1 shows an example of an input sentence annotated with POS and base-phrase information. Additionally, SRLs are used to enrich the POS and base-phrase annotation levels. Semantic roles are assigned to each predicate independently. As a consequence, the resulting annotation cannot be considered a tree and there is not an obvious hierarchy of predicates in a sentence. For example, Figure 2 shows the SRL annotation for the example in Figure 1.

[A0 He] [T intends] [A1 to donate this money to charity], but he has not decided which yet
[A0 He] intends to [T donate] [A1 this money] [A2 to charity], but he has not decided which yet
He intends to donate this money to charity, but [A0 he] has [AM-NEG not] [T decided] [A1 which] [AM-TMP yet]

Figure 2: SRL for sentence in Figure 1

Arguments of a single predicate never overlap, however in longer sentences, the occurrence of multiple verbs increases the chances that arguments of different predicates overlap, that is, the argument of a verb might contain or even coincide with the argument of another verb and depending on the verb the argument role might change. For example, in Figure 2: i) *He* is both the agent of *intend* and *donate*; ii) *this money* is the donated thing and also part of the chunk which express the intention (*to donate this money to charity*). In a different example we can see that arguments might overlap and their roles change completely depending on their target predicates (e.g in *I gave you something to eat*, *you* is the recipient of the verb *give* and the agent of the verb *eat*). For this reason, why semantic role labels are usually annotated individually in different structures, as shown in Figure 2, each annotation focusing on a single target verb. In order to convert the predicates and arguments of a sentence into a single tree, we enrich the POS-tags and base-phrase annotation as follows:

- Semantic labels are directly grafted to the base-phrase annotation whenever possible, that is, if a predicate argument coincides with a single base-phrase, the base-phrase type is specialized with the argument role. In Figure 3, the noun-phrase (NP) *the money* is specialized into *NP:A1:donate*, since that single NP is the argument A1 of *donate*.

- If a predicate argument groups multiple base-phrases, the semantic label applies to a node in a new level of the tree subsuming all these base-phrases. In Figure 3, the base-phrases *to* (PP) and *charity* (NP) are grouped by *A2:donate*.

- We add the labels sequentially from the shortest chunks to the largest ones. If two labels spanning the same number of tokens: i) overlap completely, we merge them so that no hierarchy is imposed between their targets (e.g. in Figure 3, the noun-phrase *He* is specialized into *NP:A0:donate,intend*); ii) overlap partially, we merge them so that the resulting label will compete against other labels in a different length category. If a label spanning a larger chunk overlaps partially with a label spanning a shorter chunk, or contains it, we stack them in a way that the first subsumes the second (e.g in Figure 3, *A1:intend* subsumes *VP:T:donate*, *NP:A1:donate,intend* and *A2:donate*).

- Verb phrases might get split if they contain multiple target predicates (e.g. in Figure 3, the VP *intends to donate* is split into two verb-

phrases, each specialized with its own role label).

- Finally, tags are lexicalized, that is, semantic labels are composed by their type (e.g. *A0*) and target predicate lemma (verb).

Figure 3 shows and example of how semantic labels are combined with shallow syntax in order to produce the input tree for the sentence in Figure 1. The argument *A1* of *intend* subsumes the target verb *donate* and its arguments *A1* and *A2*; *A2:donate* groups base-phrases so as to attach the preposition to the noun phrase.

Finally, following the method for syntactic trees by Venugopal and Zollmann (2006), the input trees are relaxed in order to alleviate the impact of the linguistic constraints on rule extraction. We relax trees[3] by combining any pairs of neighboring nodes. For example, *NP:A0:donate,intend+VP:T:intend* and *NP:A1:donate+A2:donate* are created for the tree in Figure 3.

4 Results

As a baseline to compare against our proposed approach (**srl**), we took a phrase-based SMT system (**pb**) built using the Moses toolkit with the same datasets and training conditions described in Section 3. The results are reported in terms of standard BLEU (Papineni et al., 2002) (and its case sensitive version, BLEU-c) and tested for statistical significance using an approximate randomization test (Riezler and Maxwell, 2005) with 100 iterations.

In addition, we included an intermediate model between these two: a hierarchical model informed with source-language base-phrase information (**chunk**). For the English-Spanish task we also built a purely hierarchical model (**hier**) using Moses and the same datasets and training conditions. For the English-German task, hierarchical models have not been shown to outperform standard phrase-based models in previous work (Koehn et al., 2010).

Table 1 shows the performance achieved for the English-Spanish translation task test set, where (**srl**) is our official submission. One can notice a significant gain in performance (up to 6% BLEU) in using tree-based models (with or without source language annotation) as opposed to using standard phrase-based models.

Model	BLEU	BLEU-c
pb	0.2429	0.2340
srl	0.2901	0.2805
hier	0.3029	0.2933
chunk	0.3034	0.2935

Table 1: English-Spanish experiments - differences between all pairs of models are statistically significant with 99% confidence, except for the pair (**hier**, **chunk**)

The purely hierarchical approach performs as well as our linguistically informed tree-based models (**chunk** and **srl**). On the one hand this finding is somewhat disappointing as we expected that tree-based models would benefit from linguistic annotation. On the other hand it shows that the linguistic annotation yields a significant reduction in the number of unnecessary productions: the linguistically informed models are much smaller than **hier** (Table 5), but perform just as well. Whether the linguistic annotation significantly helps make the productions less ambiguous or not is still a question to be addressed in further experimentation.

Table 2 shows the performance achieved for the English-German translation task test set. These results indicate that the linguistic information did not lead to any significant gains in terms of automatic metrics. An in-depth comparative analysis based on a manual inspection of the translations remains to be done.

Model	BLEU	BLEU-c
pb	0.1398	0.1360
srl	0.1381	0.1344
chunk	0.1403	0.1367

Table 2: English-German experiments - differences between pairs of models are not statistically significant

In Table 3 we also show the impact of three compound merging strategies as post-processing for en-de: i) no compound merging (**nm**), ii) frequency-based compound merging (**fb**), and iii) frequency-

[3] Using the Moses implementation *relax-parse* for SAMT 2

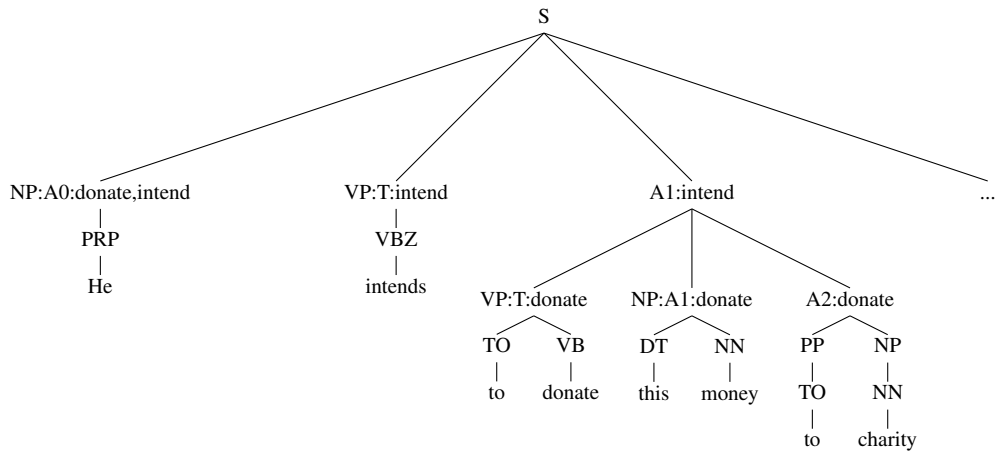

Figure 3: Tree for example in Figure 1

based compound merging constrained by POS[4] (**cfb**). Applying both frequency-based compound merging strategies (Stymne, 2009) resulted in significant improvements of nearly 0.5% in BLEU.

Model	BLEU	BLEU-c
nm	0.1334	0.1298
fb	0.1369	0.1332
cfb	0.1381	0.1344

Table 3: English-German compound merging - differences between all pairs of models are statistically significant with 99% confidence

Another somewhat disappoint result is the performance of **srl** when compared to **chunk**. We believe the main reason why the **chunk** models outperform the **srl** models is data sparsity. The semantic information, and particularly the way it was used in this paper, with lexicalized roles, led to a very sparse model. As an attempt to make the **srl** model less sparse, we tested a version of this model without lexicalizing the semantic tags, in other words, using the semantic role labels only, for example, *A1* instead of *A1:intend* in Figure 3. Table 4 shows that models with lexicalized semantic roles (*lex*) consistently outperform the alternative version (*non lex*), although the differences were only statistically significant for the en-de dataset. One reason for that may be that non-lexicalized rules do not help making the **chunk** rules less ambiguous.

Model	BLEU	BLEU-c
en-es$_{non\ lex}$	0.2891	0.2795
en-es$_{lex}$	0.2901	0.2805
en-de$_{non\ lex}$	0.1319	0.1284
en-de$_{lex}$	0.1381	0.1344

Table 4: Alternative model with non-lexicalized tags - differences are statistically significant with 99% confidence for en-de only

Table 5 shows how the additional annotation constrains the rule extraction (for the en-es dataset). The unconstrained model **hier** presents the largest rule table, followed by the **chunk** model, which is only constrained by syntactic information. The models enriched with semantic labels, both the lexicalized or non-lexicalized versions, contain a comparable number of rules. They are at least half the size of the **chunk** model and about 9 times smaller than the **hier** model. However, the number of nonterminals in the lexicalized models highlights the sparsity of such models.

Model	Rules	Nonterminals
hier	962,996,167	1
chunk	235,910,731	3,390
srl$_{non\ lex}$	92,512,493	44,095
srl$_{lex}$	117,563,878	3,350,145

Table 5: Statistics from the rule table

In order to exemplify the importance of having

[4]POS tagging was performed using the TreeTagger toolkit: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

some form of lexicalized information as part of the semantic models, Figure 4 shows two predicates which present different semantic roles, even though they have nearly the same shallow syntactic structure. In this case, unless lexicalized, rules mapping semantic roles into base-phrases become ambiguous. Besides, the same role might appear several times in the same sentence (Figure 2). In this case, if the semantic roles are not annotated with their target lemma, they bring additional confusion. Therefore, the model needs the lexical information to distinguish role deletion and reordering phenomena across predicates.

Figure 4: Different SRL for similar chunks

[NP:A0 I] [VP:T gave] [NP:A2 you] [NP:A1 a car]
[NP:A0 I] [VP:T dropped] [NP:A1 the glass] [AM-LOC [PP on] [NP the floor]]

In WMT11's official manual evaluation, our system submissions (**srl**) were ranked 10^{th} out of 15 systems in the English-Spanish task, and 18^{th} out of 22 systems participating in the English-German task. For detailed results refer to the overview paper of the Shared Translation Task of the Sixth Workshop on Machine Translation (WMT11).

5 Conclusions

We have presented an effort towards using shallow syntactic and semantic information for SMT. The model based on shallow syntactic information (chunk annotation) has significantly outperformed a baseline phrase-based model and performed as well as a hierarchical phrase-based model with a significantly smaller number of translation rules.

While annotating base-phrases with semantic labels is intuitively a promising research direction, the current model suffers from sparsity and representation issues resulting from the fact that multiple predicates share arguments within a given sentence. As a consequence, shallow semantics has not yet shown improvements with respect to the chunk-based models.

In future work, we will address the sparsity issues in the lexicalized semantic models by clustering predicates in a way that semantic roles can be specialized with semantic categories, instead of the verb lemmas.

References

Vamshi Ambati and Alon Lavie. 2008. Improving syntax driven translation models by re-structuring divergent and non-isomorphic parse tree structures. In *The Eight Conference of the Association for Machine Translation in the Americas (AMTA)*.

Kathryn Baker, Michael Bloodgood, Chris Callison-burch, Bonnie J. Dorr, Nathaniel W. Filardo, Lori Levin, Scott Miller, and Christine Piatko. 2010. Semantically-informed syntactic machine translation: A tree-grafting approach.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceeding of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *arXiv:1103.0398v1*.

Hieu Hoang and Philipp Koehn. 2010. Improved translation with source syntax labels. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 409–417.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of International Workshop on Spoken Language Translation*, pages 152 – 159.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, pages 187–193.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227.

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 *45th Annual Meeting of the Association for Computational Linguistics*.

Philipp Koehn, Barry Haddow, Philip Williams, and Hieu Hoang. 2010. More linguistic annotation for statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 115–120.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Stefan Riezler and John Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Workshop in Intrinsic and Extrinsic Evaluation Measures for MT and Summarization*.

Sara Stymne. 2009. A comparison of merging strategies for translation of german compounds. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 61–69.

Ashish Venugopal and Andreas Zollmann. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141.

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.