Improving English to Spanish Out-of-Domain Translations by Morphology Generalization and Generation

Lluís Formiga Adolfo Hernández José B. Mariño Enric Monte
Universitat Politècnica de Catalunya (UPC), Barcelona, 08034 Spain
{lluis.formiga,adolfo.hernandez,jose.marino,enric.monte}@upc.edu

Abstract

This paper presents a detailed study of a method for morphology generalization and generation to address out-of-domain translations in English-to-Spanish phrase-based MT. The paper studies whether the morphological richness of the target language causes poor quality translation when translating out-of-domain. In detail, this approach first translates into Spanish simplified forms and then predicts the final inflected forms through a morphology generation step based on shallow and deep-projected linguistic information available from both the source and target-language sentences. Obtained results highlight the importance of generalization, and therefore generation, for dealing with out-of-domain data.

1 Introduction

The problems raised when translating into richer morphology languages are well known and are being continuously studied (Popovic and Ney, 2004; Koehn and Hoang, 2007; de Gispert and Mariño, 2008; Toutanova et al., 2008; Clifton and Sarkar, 2011; Bojar and Tamchyna, 2011).

When translating from English into Spanish, inflected words make the lexicon to be very large causing a significant data sparsity problem. In addition, system output is limited only to inflected phrases available in the parallel training corpus (Bojar and Tamchyna, 2011). Hence, phrase-based SMT systems cannot generate proper inflections unless they have learned them from the appropriate phrases. That would require to have a parallel corpus containing all possible word inflections for all phrases available, which it is an unfeasible task.

Different approaches to address the morphology into SMT may be summarized in four, not mutually exclusive, categories: *i*) factored models (Koehn and Hoang, 2007), enriched input models (Avramidis and Koehn, 2008; Ueffing and Ney, 2003), segmented translation (Virpioja et al., 2007; de Gispert et al., 2009; Green and DeNero, 2012) and morphology generation (Toutanova et al., 2008; de Gispert and Mariño, 2008; Bojar and Tamchyna, 2011).

Whereas segmented translation is intended for agglutinative languages, translation into Spanish has been classically addressed either by factored models (Koehn and Hoang, 2007), enriched input scheme (Ueffing and Ney, 2003) or target language simplification plus a morphology generation as an independent step (de Gispert and Mariño, 2008). This latter approach has also been used to translate to other rich morphology languages such as Czech (Bojar and Tamchyna, 2011).

The problem of morphology sparsity becomes crucial when addressing translations out-of-domain. Under that scenario, there is a high presence of previously unseen inflected forms even though their lemma could have been learned with the training material. A typical scenario out-of-domain is based on weblog translations, which contain material based on chat, SMS or social networks text, where it is frequent the use of second person of the verbs. However, second person verb forms are scarcely populated within the typical training material (e.g. Europarl, News and United Nations). That

is due to the following reasons: *i*) text from formal acts converts the second person (*tú*) subject into *usted* formal form, which uses third person inflections and *ii*) text from news is mainly depicted in a descriptive language relegating second person to textual citations of dialogs that are a minority over all the text.

Some recent domain-adaptation work (Haddow and Koehn, 2012) has dealt implicitly with this problem using the OpenSubtitles[1] bilingual corpus that contains plenty of dialogs and therefore second person inflected Spanish forms. However, their study found drawbacks in the use of an additional corpus as training material: the improvement of the quality of the out-of-domain translations worsened the quality of in-domain translations. On the other hand, the use of an additional corpus to train specific inflected-forms language generator has not yet been addressed.

This paper presents our findings on tackling the problem to inflect out-of-domain verbs. We built a SMT system from English into simplified morphology Spanish in order to inflect the verbs as an independent postprocessing step. This strategy has been formerly applied to translate from English into Spanish with a N-gram based decoder (de Gispert and Mariño, 2008) but without dealing with out-of-domain data and neither with a factored based system (Koehn and Hoang, 2007). We analyze the most convenient features (deep vs. shallow) to perform this task, the impact of the aforementioned strategy when using different training material and different test sets. The main reason to focus the study only on the verbs is their strong impact on the translation quality (Ueffing and Ney, 2003; de Gispert and Mariño, 2008).

In section 2 we describe the architecture of the simplification plus generation strategy. In section 3 we detail the design of the generation system. In section 4 we detail the experiments performed and we discuss them in section 5. At last, we explain in section 6 the main conclusions and lines to be dealt in the future.

[1] www.opensubtitles.org

2 System architecture

The main idea of the presented strategy is to reduce the sparsity of the translation models and the perplexity of the language models by simplifying the morphology in the target language.

Spanish, as a Latin derived language, has a complex grammar. Rodríguez and Carretero (1996) enumerated the problems of Spanish morphology flexions into 7 different problems that contain verb conjugation, gender/number derivations and enclitic forms among others. As it has been mentioned, we focus on the surface forms related to Spanish verbs. Concretely we center our study to predict *i*) person and number (PN) for the Spanish verb forms and *ii*) number and gender (NG) of participles and adjectives derived from verbs, which are very common in passive forms. We implicitly deal with enclitic forms through a segmentation step based on the work by Farrús et al. (2011).

The idea is summarized in Figure 1. Spanish verb forms are replaced with their simplified form. Generalization is carried out through several steps detailed in Table 1. The Spanish POS tags are given in Parole format[2] that includes information about the type, mode, tense, person, number and gender of the verb. First, we concatenate the POS tag to the lemma of the verb. For example, the inflected form *puede* is transformed into VMIP3S0[poder], which indicates that the lemma of Main Verb poder is inflected to the Indicative Present Third Person Singular form. Next, we generalize the person, number and gender of the verb to the following variables: p for person, n for number and g for gender. Under this generalization, the simplified form keeps information of verb type ('VM' $\rightarrow$ main verb), mode and tense ('IP' $\rightarrow$ indicative, present), while 'p' and 'n' represent any person and number once generalized (from 3rd person singular). It is important to highlight that we do not perform just a simple *lemmatization* as we also keep the information about the type, mode and tense of the verb.

After simplifying the corpus we can build the models following the standard procedures explained in section 4.1. Note that the tuning of the system is performed with the simplified reference of the development texts.

[2] http://www.lsi.upc.edu/ nlp/tools/parole-eng.html

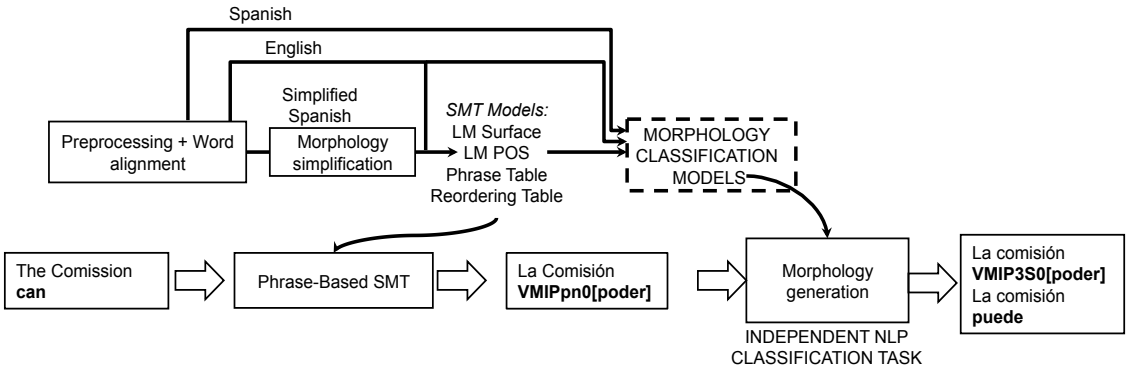

Figure 1: Flow diagram of the training of simplified morphology translation models.

Type	Text
PLAIN TARGET:	la Comisión **puede** llegar a paralizar el programa
Lemma+PoS	la Comisión **VMIP3S0[poder]** llegar a paralizar el programa
Lemma+PoS Generalized:	la Comisión **VMIPpn0[poder]** llegar a paralizar el programa

Table 1: Example of morphology generalization steps taken for Spanish verbs.

At this point, the translation process may be independently evaluated if test references are also simplified. This evaluation provides oracles for the generation step. That is, the maximum gain to be obtained under a perfect generation system.

Finally, morphology prediction system is designed independently as it is explained in section 3. The generation system predicts the correct verb morphology for the given context both in the source and the target sentence. Once the morphology is predicted, the verb is inflected with a verb conjugator.

The presented strategy has two clear benefits: *i*) it makes clear the real impact of morphology generalization by providing an oracle for the studied scenarios and *ii*) decouples the morphology generation system from the actual SMT pipeline making it feasible to be trained with small or noisy out-of-domain corpora without having a strong negative impact into the decoder pipeline (Haddow and Koehn, 2012).

However, any bilingual corpora used to train the generation system has to be correctly aligned in order to perform a correct extraction of the features. In that sense it is useful to reuse the already trained SMT (e.g. GIZA) alignment models as they are built from larger collections of data.

3 Design of the Generation System

The generation system is addressed as a multiclass classification problem. We separate the prediction in two independent tasks: *i*) person and number and *ii*) number and gender. The reason of the separation is the fact that in Spanish there are not verb forms where the person, number and gender have to be predicted at the same time. Thus, the forms other than participle involve decisions only based in person and number while the participle forms involve only number and gender. Thus, we train two independent multiclass classifiers: *i)* a person and number classifier involving 6 output classes (1st, 2nd and 3rd person either in Singular or Plural) and *ii)* a number and gender classifier involving 4 output classes (Male and Female either in Singular or Plural). We provide the one-best decision of the decoder as the input to the generation system along with its related tokenized source sentence and its alignment. It is important to highlight that the decoder has to be able to provide the source-translation alignment at word level.

3.1 Relevant Features

A set of linguistic features is extracted for each generalized verb found in the target sentence. These features include simple shallow information around the verb and might include deep information such as projected dependency constituents or semantic role labels.

For the shallow feature extraction, the features are extracted with simple neighborhood functions that look the words, POS tags and the morphology in and around the verb in both the source and target side. These features are: *i*) Context words and

their POS for both the source and target verbs. *ii*) The composed verb phrase and its POS (e.g. it has not already been saved). The verb phrase is detected through a WFST acceptor. We also consider mixed word/POS source verb sequences (e.g. PRP has not already been VB). *iii*) Presence of a passive voice on the source. *iv*) Sequence of named entities (and their conjugations) before the source and target verbs: (e.g, John, Mary and Peter). *v*) Reflexive pronoun after the source and target verbs. *vi*) Pronoun before the source verb or whether it has POS indicating 3S (VBZ) or not3S (not VBZ) conjugation. *vii*) Pronoun before the target verb (yo, tú...). *viii*) Simplified form of the target verb simplifying also its mode and mode and tense. *ix*) Abstract pattern of the verb noting whether it is a auxiliary *haber* plus participle or simply a participle (mainly used as adjective).

For the deep features, first we perform semantic role labeling and dependency parsing of the source sentence through the Semantic parser of Lund University [3] and then we project this information to the target side using the alignment. In case of alignment to multiple words, we use the lexical model probabilities to decide the target word that corresponds to the source dependency. In total we use 310 different deep features such as: pA (parent agent), cSBJ (child subject), cOB (child object), pNMOD (parent modifier), pA1_pos (POS of the parent agent 1) among others. The most important learned features are detailed in Section 4.4.

3.2 Classifier framework

The generation system is implemented by means of classification models that predict the person, number and gender from the extracted features. Typical algorithms to deal with this task are Conditional Random Fields (McCallum and Li, 2003), MaxEnt classifiers (Della Pietra et al., 1997) or Support Vector Machines (Platt et al., 2000). All of them usually represent the set of features as a binary array.

We discard CRFs because the prediction case described in this paper does not suit as a structured/sequential problem as we only focus on predicting verb forms and usually they don't influence each other within the sentence and therefore each of them becomes a root constituent itself.

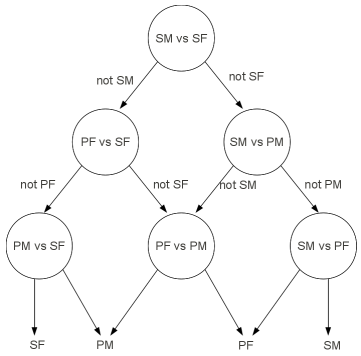

Figure 2: Decision DAG to find the best class out of four classes related to gender and number

We have chosen SVMs instead of MaxEnt because the feature vectors are high-dimensional. Concretely, the binary vector size is 380k for the shallow features and 755k for the vectors that combine shallow and deep features. Therefore, SVM approximates the decision boundary by means of support vectors, which allow curvature in the feature space when it is high dimensional. This was confirmed in some preliminary experiments where we found better performance and that the size of the support vectors was about the 5% with respect to the total training database. On the other hand the MaxEnt classifier is based on simple hyperplanes, which assumes that the underlying boundary between classes is linear. In addition, the MaxEnt model assumes that the distribution of the the dot product between the feature vector and the set of weights of the classifier, which in the model is reflected by the use of an exponential nonlinearity. This assumption is rather limited and might not be correct.

Among the different multiclass SVM approaches, we have implemented the generation system by Decision Directed Acyclic Graphs (DDAG) (Platt et al., 2000) composed of binary SVM classifiers. A DDAG combines many two-class classifiers into a multiclassification task. The description of the structure is as follows: For an N-class problem, the DDAG contains N(N-1)/2 nodes, one for each pair of classes (one-vs-one classifier). A DAGSVM algorithm is proposed by Platt et al. (2000). An example of a structure of the DDAG is shown in Figure 2.

The classifiers can be ordered following different

[3] http://nlp.cs.lth.se/software/

criteria such as the misclassification rate, the balance between samples of each class or the most reliable decision taken by the classifiers. In this paper we follow the latter criteria: After processing the features by all classifiers simultaneously, the most consistent decision from all binary classifiers is taken in first place, afterwards the second best is considered and so on, until the final class is answered by the binary decisions. The experiments are explained in section 4.4.

4 Experiments

The experiments were carried out in three distinct stages. First, we have analyzed the impact of morphological generalization into the decoder models both with texts of the training-domain and text out-of-domain. Then, we studied the generation system accuracy with fluent text sets and finally, we have studied the overall improvement achieved by the whole strategy under the different scenarios.

4.1 Baseline systems

Corpus		Sent.	Words	Vocab.	avg.len.
EPPS	Eng	1.90 M	49.40 M	124.03 k	26.05
	Spa		52.66 M	154.67 k	27.28
News.Com	Eng	0.15 M	3.73 M	62.70 k	24.20
	Spa		4.33 M	73.97 k	28.09
UN	Eng	8.38 M	205.68 M	575.04 k	24.54
	Spa		239.40 M	598.54 k	28.56

(a) Parallel

Corpus	Sent.	Words	Vocab.
EPPS	2.12 M	61.97 M	174.92 k
News.Com.	0.18 M	5.24 M	81.56 k
UN	11.20 M	372.21 M	725.73 k
News.07	0.05 M	1.33 M	64.10 k
News.08	1.71 M	49.97 M	377.56 k
News.09	1.07 M	30.57 M	287.81 k
News.10	0.69 M	19.58 M	226.76 k
News.11	5.11 M	151.06 M	668.63 k

(b) Monolingual

Table 2: Details of different corpora used for training the models. The counts are computed before generalization.

We based our experiments under the framework of a factored decoder (Moses – Koehn and Hoang (2007)). Concretely, we translate the source words into target words plus their POS tags (Factored Moses from 0 to 0,2) using two separate language models for improving the fluency of the output. We did the alignment with stems through mGIZA (Gao and Vogel, 2008). We used the material from WMT12 (Callison-Burch et al., 2012) MT Shared Task for training. We used the Freeling analyzer (Padró et al., 2010) to tokenize, lemmatize and POS-tag both sides of the corpus (English and Spanish). In the same way we use the Freeling libraries in order to conjugate the verbs. We trained the language models (LM) with the SRILM Toolkit (Stolcke, 2002) at 5-gram level for words and 7-gram level for POS-tags.

In order to study the impact of the morphology at different training levels we have considered two different scenarios: First, we train a system only with texts from the European Parliament being a limited resource scenario, hereafter EPPS, consisting of small-sized corpora. Secondly, we consider a state-of-the-art scenario, hereafter WMT12, using all the material available. Corpus details are given in table 2. Weights have been tuned according to the development material of WMT'12 (7567 news sentences from 2008 to 2010). The news material for the years 2011 and 2012 has been left aside for testing purposes as explained later.

All these steps were performed identically for both the baseline and simplified verb forms decoders. Note that for the latter, the POS factor is also simplified. In addition, we needed also to simplify the development texts for tuning the system.

4.2 Test scenarios

We set different evaluation test sets: news tests from WMT11 and WMT12 (Callison-Burch et al., 2012) for in-domain evaluation and weblog translations from the FAUST project (Pighin et al., 2012) for the out-of-domain. The news sets from WMT consist of 3003 human revised translations each. They will be referred as n11 and n12 in this paper. Regarding the weblog translations we considered 998 translation requests in English into Spanish submitted to Softissimo's online translation portal[4]. Two independent human translators had corrected the most obvious typos and provided reference translations into Spanish for all of them along with the clean versions of

[4]http://www.reverso.net

the input requests. Thus, we consider four different test sets from this material:

i) *Weblog Raw (wr)* The noisy weblog input. It contains misspellings, slang and other input noise typical from chats, forums, etc. These translations are evaluated with their correspondent reference provided by each translators (two references).

ii) *Weblog Clean$_i$ (w0 and w1)* The cleaned version of the input text provided by each translator on the source side. Cleaned versions may differ due to the interpretation of the translators (e.g. If you dont like to chat → If you don't like chatting — If you don't want to chat).

iii) *Weblog Clean0.1 (w0.w1)* In that case we mix up the criteria of the different translators. In that case the cleaned versions are concatenated (making up a set of 1,996 sentences) and evaluated with their respective translations (two references).

4.3 Impact of morphology generalization into the Decoder

We analyzed the effect of the morphology generalization into the decoder's models across two different aspects. First, we analyzed to what extent the morphology generalization reduces the perplexity of the language models built upon words and POS tags. Secondly, we analyzed the downsizing of the sparsity within the Moses lexical models.

Results of the perplexity and sparsity reduction are detailed in table 3. The EPPS results detail the reduction within the constrained decoder and the WMT12 ones detail the reduction within the fully-trained decoder. In general terms, word level perplexities are reduced by a 6-7% when working with formal News data (in-domain) and by a 12-17% when working with weblog data. We observed that perplexity reduction is relatively more important for the constrained system. For the POS Language Models we observed less margin of reduction for the in-domain News sets (3-6%) and similar results for the weblog dataset (11.5-18%). With respect to the lexical models, we observed a reduction of the Spanish unique entries of the model. For the constrained system (EPPS) the entries are reduced from 164.13k to 140.10k and for the fully trained (WMT12) system the entries are reduced from 660.59k to 626.36k. The ratios of the lexical models show that the sparsity is clearly defined in the constrained system while it becomes balanced with a larger training corpus. In the latter case the generalization causes a negative sparsity relation.

EPPS	**Base**	**Simp.**	**%**
n11	291.63	270.61	-7.21
n12	288.66	267.19	-7.44
w0	944.18	790.46	-16.28
w1	1076.28	910.67	-15.39
WMT12	**Base**	**Simp.**	**%**
n11	186.04	174.74	-6.07
n12	172.65	162.29	-6.00
w0	613.38	533.73	-12.99
w1	645.00	563.27	-12.67

(a) Word perplexity

EPPS	**Base**	**Simp.**	**%**
n11	15.21	14.31	-5.92
n12	15.84	14.87	-6.12
w0	43.33	35.46	-18.16
w1	50.12	41.63	-16.94
WMT12	**Base**	**Simp.**	**%**
n11	12.74	12.33	-3.22
n12	13.1	12.47	-4.81
w0	30.07	26.33	-12.44
w1	33	29.21	-11.48

(b) PoS perplexity

EPPS	**English**	**Spanish**	**Ratio**
Base	124.06k	164.13k	1.32
Simp.		140.10k	1.13
WMT12			
Base	658.67k	660.59k	1.00
Simp.		626.36k	0.95

(c) Lexical Entries

Table 3: Evaluation of perplexity and lexical entries reduction obtained by the morphology generalization strategy.

4.4 Generation System

After analyzing the impact of the generalization strategy into the decoder models, we evaluated the DDAG accuracy to predict the morphology of the verb forms.

Previous studies (de Gispert and Mariño, 2008) detailed that the learning curve for predicting verb forms stabilized with 300,000 verb samples for PN and 150,000 verb samples NG. As the purpose of this paper is to analyze the suitability of

DDAG accuracy		Test sets							AVG
Person and Number		***wmt12***	***Sub***	***n08-10***	***w0***	***w0.w1***	***w1***	***wr***	
Shallow	*wmt12*	86.59	68.39	84.41	73.89	74.56	74.77	71.87	76.35
	wmt12+Sub	85.71	80.30	84.76	83.41	84.39	84.46	81.52	83.51
	Subtitles	80.75	81.87	82.73	84.32	84.57	84.28	82.48	83.00
Shallow+Dep	*wmt12*	**87.67**	68.45	84.93	73.80	74.24	74.22	71.96	76.47
	wmt12+Sub	86.78	80.50	**85.44**	84.68	84.75	84.19	82.02	**84.05**
	Subtitles	81.81	**82.00**	83.21	**85.04**	**84.98**	**84.92**	**82.70**	83.52
Number and Gender									
Shallow	*wmt12*	88.09	86.25	**84.07**	79.82	80.74	80.77	80.95	82.96
	wmt12+Sub	86.63	90.06	83.93	**83.77**	**84.20**	**84.62**	**83.98**	**85.31**
	Subtitles	80.46	88.06	82.79	81.14	81.39	81.20	81.39	82.35
Shallow+Dep	*wmt12*	**88.60**	86.49	84.00	81.58	80.52	81.20	80.74	83.30
	wmt12+Sub	87.16	**90.49**	83.71	**83.77**	83.55	83.76	83.12	85.08
	Subtitles	80.82	88.09	82.06	82.89	82.90	82.91	82.68	83.19

Table 5: Accuracy scores achieved by the DDAG learner trained with different clean and aligned corpus (wmt12, Subtitles and combined) and different feature sets (Shallow and Shallow+Dependencies). The best results are depicted in bold.

		PN		NG	
		Train	**Test**	**Train**	**Test**
WMT12		300k	189k	150k	40k
Subtitles		300k	82k	30k	7k
Combined	**WMT12**	150k	339k	120k	70k
	Subtitles	150k	232k	30k	7k
	Total	300k	570k	150k	77k

Table 4: Details of the number of verbs per corpora and task used for training the generation system. PN stands for Person and Number and NG for Number and Gender.

morphology-generalization strategy when addressing out-of-domain translations, we did not consider the study a new learning curve

We trained the generation system with clean and fluent corpora (not MT-output). Details of the different corpora studied are depicted in table 4.

First, we trained as a baseline generation system with the same corpora of WMT12. We homogeneously sampled 300,000 sentences from the parallel corpus with 678k verbs. We used 450,000 verbs for training the generation system (300,000 for person and number (PN) and 150,000 for number and gender (NG)) setting aside 228k verbs (188 for PN and 40k for NG) for testing purposes.

We coped with second person morphology (*tú* / *vosotros*) with the use of OpenSubtitles corpora as training material, which contains plenty of dialogs. In that case we needed to align the sentences. We performed all the steps of mGIZA starting from the previously trained WMT12 models.

We used the OpenSubtitles corpora in two different ways: entirely or partially combined with the WMT12 corpora. However, the Subtitles corpora does not have enough verb forms for training the number and gender system, causing a smaller size of the training set for the standalone system and not allowing an equal contribution (50%) for the combined version.

w0.w1	Stats			
PN	**Precision**	**Recall**	**Specifity**	**F1**
1S	0.93	0.88	0.98	0.45
2S	0.80	0.80	0.97	0.40
3S	0.82	0.92	0.86	0.44
1P	0.89	0.79	1.00	0.42
2P	0.00	0.00	1.00	0.00
3P	0.82	0.67	0.98	0.37
NG				
SM	0.86	0.94	0.73	0.45
SF	0.78	0.63	0.96	0.35
PM	0.80	0.70	0.98	0.37
PF	0.84	0.73	0.99	0.39

Table 6: Classification scores for the best accuracy configurations.

We also tested the prediction task in sets other than the verbs left apart from the training data. Concretely, we used the development material of WMT12 (n08-10) and the weblog test data.

Results are shown in tables 5. Regarding the feature sets used, as explained on section 3.1, we analyzed the accuracy both with shallow features and combining them with deep projected features (*Shallow+Dep*) based on syntactic and semantic dependencies. We also analyzed the precision, recall and F1 scores for each class for the *w0.w1* test set (Table 6). These results are from the best configurations achieved (PN: *Shallow+Dep* trained only with Subtitles and NG: *Shallow* trained with combined sets (WMT12+Sub)).

Results to predict person and number indicate that models trained with only subtitles yield the best accuracies for weblog data, whereas the models trained with the *WMT12+Sub* combined set yield the best results for the News domain. In addition, we observed that the best results are obtained with the help of the deep features indicating that they are important for the prediction task.

However, deep features do not help in the prediction of number and gender for the weblog and News test sets. With respect to the training material, the best results are achieved by the combined training set *WMT12+Sub* for the weblog tests and by the standalone WMT12 set for the News test set. This behavior is explained by the small amount of number and gender samples in the subtitles set.

Consequently, we analyzed the most important features from the DDAG-SVM models, i.e. those features with a significant weight values in the support vectors of the classifiers. Regarding the PN classifiers, we found that the Shallow features were among the 9 most important features of the PN models. Dependency features were less important being the POS, surface and lemma of the subject the 10th, 13th and 16th most important features respectively. Predicate features had a minimal presence in the models being the POS of the APP0 the 24rd most important feature. As presumed, for the NG classifiers the impact of the deep features was less important. In that case the POS of the NMOD and PMOD were in the 14th and 17th positions respectively and the POS of A1 the 18th most important feature.

With respect to the correctness of the classifiers per class (Table 6), we observed that 1P and SM classes are the ones with the highest F1 score. However, 2P class cannot be predicted due to its small presence ($\approx$ 0.6%) in both training and testing sets. When analyzing the results in detail, we found considerable confusions between 3P-3S, 2S-3S, and SM-SF. This latter case is caused by the presence of female proper nouns that the system is not able to classify accordingly (e.g. *Tymoshenko*) and therefore assigns them to the majority class (SM). All the F1 scores are around 0.35 and 0.45 per class, with the exception of 2P that can not be predicted properly.

4.5 Translation

Before analyzing the improvement of the strategy as a whole, we made an oracle analysis without the generation system. In that case, we evaluated the oracle translations by simplifying the reference translations and comparing them to the output of the simplified models. We detail the BLEU oracles in table 7. For the constrained system we observed a potential improvement between 0.5 to 0.7 BLEU points for the News sets and an improvement from 1 to 1.3 BLEU points for weblog datasets. For the full trained system we observed a similar improvement for the News sets (between 0.5 and 0.7 BLEU points) but a better improvement, between 2 and 3 BLEU points, for the out-of-domain weblog data. These oracles demonstrate the potential of morphology generalization as a good strategy for dealing with out-of-domain data.

After analyzing the oracles we studied the overall translation performance of the strategy. We analyzed the results with BLEU and METEOR (Denkowski and Lavie, 2011). However, METEOR properties of synonymy and paraphrasing did not make it suitable for evaluating the oracles for the simplified references. In addition, table 7 details the results for the full generation strategy. In general terms, we observe better improvements for the weblog (out-of-domain) data than for the News data. For the constrained system, weblog test sets improve by 0.55 BLEU/0.20 METEOR points while News test sets only improve 0.25 BLEU/0.14 METEOR points. For the fully trained system, the out-of-domain improvement is 1.49 BLEU/1.27 METEOR points in average and the News (in-domain) achieve an improvement of 0.62/0.56 METEOR points. These results are discussed next.

BLEU-EPPS		Test sets						AVG
Method	***Train***	***w0***	***w0.w1***	***w1***	***wr***	***n11***	***n12***	
Baseline	–	26.91	32.86	25.86	28.94	28.58	28.36	28.59
Oracle	–	27.97	34.17	27.01	30.06	29.35	28.87	29.57
Shallow	*wmt12*	26.87	32.82	25.83	28.85	**28.98**	**28.46**	28.64
	wmt12+Sub	**27.53**	**33.53**	**26.42**	**29.3**	28.92	**28.46**	**29.03**
	Subtitles	27.41	33.4	26.34	29.19	28.83	28.37	28.92
Shallow+Dep	*wmt12*	26.95	32.88	25.85	28.92	28.96	28.45	28.67
	wmt12+Sub	27.49	33.47	26.36	29.24	28.94	**28.46**	28.99
	Subtitles	27.38	33.39	26.34	29.19	28.86	28.39	28.93
METEOR-EPPS								
Baseline	–	52.46	55.89	52.32	**52.55**	52.62	52.67	53.08
Shallow	*wmt12*	52.36	55.89	52.28	52.29	**52.87**	**52.71**	53.07
	wmt12+Sub	**52.68**	**56.23**	**52.60**	52.51	52.85	52.70	**53.26**
	Subtitles	52.63	56.18	52.53	52.46	52.78	52.62	53.20
Shallow+Dep	*wmt12*	52.33	55.89	52.29	52.33	52.86	52.70	53.07
	wmt12+Sub	52.64	56.19	52.56	52.45	52.86	52.70	53.24
	Subtitles	52.64	56.17	52.52	52.48	52.81	52.66	53.21
BLEU-WMT12								
Baseline	–	29.07	36.02	27.92	31.81	32.62	**33.01**	31.74
Oracle	–	31.12	39.01	30.63	34.16	33.38	33.49	33.63
Shallow	*wmt12*	29.82	37.31	29.24	32.82	32.87	32.98	32.51
	wmt12+Sub	**30.59**	**38.17**	**29.94**	**33.28**	32.87	32.99	**32.97**
	Subtitles	30.43	37.92	29.78	33.12	32.77	32.87	32.82
Shallow+Dep	*wmt12*	29.87	37.35	29.23	32.82	**32.91**	32.99	32.53
	wmt12+Sub	30.55	38.09	29.9	33.26	32.89	**33.01**	32.95
	Subtitles	30.48	38.03	29.89	33.21	32.77	32.87	32.88
METEOR-WMT12								
Baseline	–	53.20	56.88	53.19	53.36	55.19	55.64	54.58
Shallow	*wmt12*	54.32	58.15	54.31	54.36	55.53	55.70	55.40
	wmt12+Sub	**54.70**	**58.58**	54.69	**54.62**	55.51	55.69	**55.63**
	Subtitles	54.61	58.45	54.59	54.53	55.44	55.62	55.54
Shallow+Dep	*wmt12*	54.27	58.14	54.31	54.35	**55.55**	**55.71**	55.39
	wmt12+Sub	54.67	58.57	**54.70**	54.61	55.53	**55.71**	**55.63**
	Subtitles	54.60	58.47	54.61	54.58	55.47	55.62	55.56

Table 7: Evaluation scores for English-Spanish translations considering Baseline, Oracle and Morphology Generation configurations. The best results are depicted in bold.

5 Discussion

The comparison of the different experiments show that a better improvement of the language models perplexity do not lead to a better improvement into the oracles obtained. Concretely, the EPPS constrained language models achieved a higher improvement with respect to the perplexities, whereas the fully trained WMT12 decoder achieved better improvement oracles. These results point the importance of the morphology generalization to the phrase-based and lexical models other than the language models.

In addition, when considering the full strategy the non-constrained system (WMT12) achieves higher improvements compared to the constrained decoder in most of the metrics. The constrained decoder provides a less fluent translation (and more noisy) compared to the fully trained decoder. Consequently, the morphology prediction task becomes more difficult for the constrained scenario due to the high presence of noise in the context of the generalized verbs. The noise presence into the MT-output also explains why the deep features do not help to obtain better translations. The main difference between the accuracy and translation experiments is the typology of the text where the prediction takes place. Whereas the accuracy experiments are performed with human references the generation system has to deal with the

decoder output, which is noisy and less fluent, making the shallow features more robust. Thus, the strategy becomes more relevant when a decoder of better quality is available because a more fluent MT-output eases the task of morphology prediction.

The combined training set (*wmt12+Sub*) achieves the most stable improvement across all the metrics and trained scenarios. The WMT12 generation system worsens the baseline results, making the Subtitles corpus a crucial part to be combined into the training material in order to achieve a high improvement for the fully trained system due to, among other reasons, the lack of second person inflected forms into the training material.

We conducted a posterior analysis of the cases when the generation system worsened the oracle. In that case we found that in the 25% of these cases the generation was correctly performed but there was a change of the subject between the reference and the output. For example, the English phrase "Good people are willing" translated as "*Las buenas personas están*" has a worse score than "*Las buenas personas está*" with the reference "*La gente buena está*". In that example the metric penalizes the good agreement instead of the verb correspondence with the reference, which obviously it is not correct.

6 Conclusions and Future Work

This paper presents a strategy based on morphology generalization as a good method to deal with out-of-domain translations, whereas it provides stability to in-domain translations. The experiments point the morphological sparseness as a crucial issue to deal when performing domain adaptation in SMT into richer languages along with language model perplexity.

In addition, we have shown that training morphology generation systems with the help of noisy data (OpenSubtitles) might help to obtain a better translation without compromising the quality of the models. Morphology generation systems might be trained with a relatively small amount of parallel data compared to standard SMT training corpora.

We have also shown the importance of projected deep features in order to predict the correct verb morphology under clean and fluent text. However, the projection of deep features is sensitive to the fluency of the sentence making them unreliable when they are applied to noisy MT-output.

Also we have shown that the morphology generation system becomes more relevant with high quality MT systems because their output is more fluent, making the shallow and deep features more reliable to guide the classifier.

Future plans include providing a n-best list or a lattice to the generation system to expand its search. We also work on the study of the projection heuristics in order to make the deep features less sensitive to the MT-output noise. Finally, we want to expand our study to the generalization of common nouns, function words and adjectives. In this case we should study the suitability of sequential learning frameworks such as CRF or probabilistic graphical models (PGM).

Acknoweledgments

We would like to thank the anonymous reviewers for their insightful comments. We also want to thank Daniele Pighin for his valuable advice. This research has been partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 247762 (FAUST project, FP7-ICT-2009-4-247762) and by the Spanish Government (Buceador, TEC2009-14094-C04-01)

References

E. Avramidis and P. Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. *Proc. of ACL-08: HLT*, pages 763–770.

O. Bojar and A. Tamchyna. 2011. Forms Wanted: Training SMT on Monolingual Data. Workshop of Machine Translation and Morphologically-Rich Languages., January.

C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proc. of the 7th Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. ACL.

A. Clifton and A. Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proc. of the 49th Annual Meeting of the ACL-HLT. Portland, OR, USA*.

A. de Gispert and J. Mariño. 2008. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*, 50(11-12):1034–1046.

A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne. 2009. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proc. of Human Language Technologies: The 2009 Annual Conference of the NAACL, Short Papers*, pages 73–76, Stroudsburg, PA, USA.

S. Della Pietra, V. Della Pietra, and J. Lafferty. 1997. Inducing features of random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(4):380–393.

M. Denkowski and A. Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proc. of the 6th Workshop on Statistical Machine Translation*, pages 85–91. Association for Computational Linguistics.

M. Farrús, M. R. Costa-jussà, J. B. Mariño, M. Poch, A. Hernández, C. A. Henríquez Q., and J. A. R. Fonollosa. 2011. Overcoming statistical machine translation limitations: error analysis. *Language Resources and Evaluation*, 45(2):165–179, May.

Q. Gao and S. Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. ACL.

S. Green and J. DeNero. 2012. A class-based agreement model for generating accurately inflected translations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–155, Jeju Island, Korea, July. Association for Computational Linguistics.

B. Haddow and P. Koehn. 2012. Analysing the effect of out-of-domain data on smt systems. In *Proc. of the 7th Workshop on Statistical Machine Translation*, pages 422–432, Montréal, Canada. ACL.

P. Koehn and H. Hoang. 2007. Factored translation models. In *Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic, June. ACL.

A. McCallum and W. Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proc. of the 7th conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

Ll. Padró, M. Collado, S. Reese, M. Lloberes, and I. Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proc. of 7th Language Resources and Evaluation Conference (LREC 2010)*, La Valletta, MALTA, May. ELRA.

D. Pighin, Ll. Màrquez, and Ll. Formiga. 2012. The faust corpus of adequacy assessments for real-world machine translation output. In *Proc. of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

J. Platt, N. Cristianini, and J. Shawe-taylor. 2000. Large margin DAGs for multiclass classification. In *Advances in Neural Information Processing Systems*, pages 547–553. MIT Press.

M. Popovic and H. Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC'04*, pages 1585–1588, May.

S. Rodríguez and J. Carretero. 1996. A formal approach to spanish morphology: the coi tools. *Procesamiento del Lenguaje Natural*, 19:119.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. of the ICSLP*, pages 311–318, Denver, Colorado, September.

K. Toutanova, H. Suzuki, and A. Ruopp. 2008. Applying morphology generation models to machine translation. In *Proc. of ACL-08: HLT*, pages 514–522, Columbus, Ohio, June. ACL.

N. Ueffing and H. Ney. 2003. Using pos information for statistical machine translation into morphologically rich languages. In *Proc. of the 10th conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 347–354, Stroudsburg, PA, USA. ACL.

S. Virpioja, J.J. Väyrynen, M. Creutz, and M. Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Machine Translation Summit XI*, 2007:491–498.