2012 AMTA 20 Years

The Tenth Biennial Conference of the
Association for Machine Translation in the Americas

Monolingual Machine Translation

Tsuyoshi Okita
Dublin City University

Artem Sokolov
LIMSI

Taro Watanabe
National Institute of Information
and Communications Technology

SAN DIEGO, CA
OCTOBER 28- NOVEMBER 1, 2012

Proceedings of the

Monolingual Machine Translation-2012 Workshop

Collocated with the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2012)

Edited by

Tsuyoshi Okita

and

Artem Sokolov

and

Taro Watanabe

1 November 2011

San Diego, USA

Preface to the MONOMT-2012 Workshop Proceedings

On behalf of the program committee of the Monolingual Machine Translation workshop (MONOMT-2012), it is our pleasure to present you this proceedings. The MONOMT-2012 workshop was held on November 1 2012, co-located with the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012) that took place from October 29 to 31, 2012 in San Diego, USA.

This workshop was the first attempt to showcase various monolingual machine translation subtasks which we frequently encounter in the recent MT research. Such subtasks include MT for morphologically rich languages [Bojar, 08], system combination strategy [Matusov et al., 05], statistical post-editing [Dugast et al., 07; Simard et al., 07], domain adaptation [Daume III, 07], two-step approach to a long-range reordering strategy (SVO and SOV) [Isozaki et al., 10], MERT process [Arun et al., 10], and translation memory and MT integration strategy [Ma et al., 11]. If we permit human-aided translation instead of MT, such as error identification and voting with independent monolingual crowdsources [Hu et al., 11] and monolingual Machine Translator [Koehn, 10], this would further enlarge the area of monolingual translation subtasks. With this showcasing, our intention was to preferably go further to obtain algorithmic tools: the modest goal would be in the form of monolingual MT tools such as MBR decoding, monolingual word alignment (based on TER and METEOR), language model learnt by its representation of data, and machine learning strategies, while the ultimate goal would be to build a general-purpose device *Monolingual Machine Translation system*. Note that we did not intend to replace the current available SMT systems, such systems would complement them.

We received 8 submissions, of which 3 were accepted as long presentations while 4 were accepted as short presentations. Submissions came from different countries: China, Czech Republic / Indonesia, Germany, India, Japan, Spain, UK and US. All the papers concerned various aspects of monolingual MT subtasks. Three papers proposed interesting ways to use monolingual resources via automatically created multiple references, domain clustering using huge web crawled data, and morphology prediction system. Regretfully, no paper concerned *generic monolingual MT tools*, probably because of the short preparation period. In this sense, our principal goal largely remains untouched until, we hope, near future; although some of our invited speakers might kindly mention them in their talks. We would like to thank invited speakers, who accepted to give a talk at this workshop: Prof. Marcello Federico, Prof. Philipp Koehn, Prof. Qun Liu, Dr. Evgeny Matusov, Dr. Taro Watanabe (At the time of writing, this list was not complete).

We note that this workshop has close links to two previous workshops: MTML 2011 workshop and ML4HMT 2011 workshop. An MTML 2011 workshop (Machine Translation and Morphologically-Rich Languages, Research Workshop of the Israel Science Foundation) was held in January 2011 in Haifa. This workshop concerned machine translation from morphologically poor into morphologically rich languages. Among its presentations, there were two that had mentioned a two-step approach for MT of morphologically rich languages. An ML4HMT 2011 workshop (Machine Learning Techniques to Optimizing the Division of Labour in Hybrid MT) was held in November 2011 in Barcelona. This workshop focused on incorporating of semantic / syntactic meta knowledge into system combination strategy. Typically, such strategy handles only monolingual MT and

we regard this as our starting point.

This workshop would not be possible without the valued help of the AMTA General Chair Prof. Alon Lavie and the AMTA workshop organizer Dr. Désillets Alain. We particularly thank Prof. Lavie for inspiring us to organize this workshop through the above two workshops. We would also like to thank all the members of the program committee half of them overlapping with the above two workshops, who have dedicated their valuable time to review the papers. Finally, we gratefully acknowledge the sponsorship of Prof. Josef van Genabith, CNGL / DCU (Center for Next Generation Localisation / Dublin City University).

1st November 2012
Tsuyoshi Okita, Artem Sokolov, and Taro Watanabe
MONOMT-2012 Workshop

Workshop Chairs

Tsuyoshi Okita (Dublin City University, Ireland)
Artem Sokolov (LIMSI, France)
Taro Watanabe (National Institute of Information and Communications Technology, Japan)

Program Committee

Bogdan Babych (University of Leeds, UK)
Loic Barrault (LIUM, Universite du Maine, France)
Nicola Bertoldi (FBK, Italy)
Ergun Bicici (CNGL, Dublin City University, Ireland)
Ondrej Bojar (Charles University, Czech)
Boxing Chen (NRC Institute for Information Technology, Canada)
Trevor Cohn (University of Sheffield, UK)
Marta Ruiz Costa-jussa (Barcelona Media, Spain)
Josep M. Crego (SYSTRAN, France)
John DeNero (Google, USA)
Jinhua Du (Xi'an University of Technology, China)
Kevin Duh (Nara Institute of Science and Technology, Japan)
Chris Dyer (CMU, USA)
Christian Federmann (DFKI, Germany)
Yvette Graham (University of Melbourne, Australia)
Barry Haddow (University of Edinburgh, UK)
Xiadong He (Microsoft, USA)
Jagadeesh Jagarlamudi (University of Maryland, USA)
Jie Jiang (Applied Language Solutions, UK)
Philipp Koehn (University of Edinburgh, UK)
Shankar Kumar (Google, USA)
Alon Lavie (CMU, USA)
Yanjun Ma (Baidu, China)
Aurelien Max (LIMSI, University Paris Sud, France)
Maite Melero (Barcelona Media, Spain)
Philip Resnik (University of Maryland, USA)
Stefan Riezler (University of Heidelberg, Germany)
Lucia Specia (University of Sheffield, UK)
Marco Turchi (JRC, Italy)
Antal van den Bosch (Radboud University Nijmegen, Netherlands)
Xianchao Wu (Baidu, Japan)
Dekai Wu (HKUST, Hong Kong)
Francois Yvon (LIMSI, University Paris Sud, France)

Table of Contents

Long Presentations

Short Presentations

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

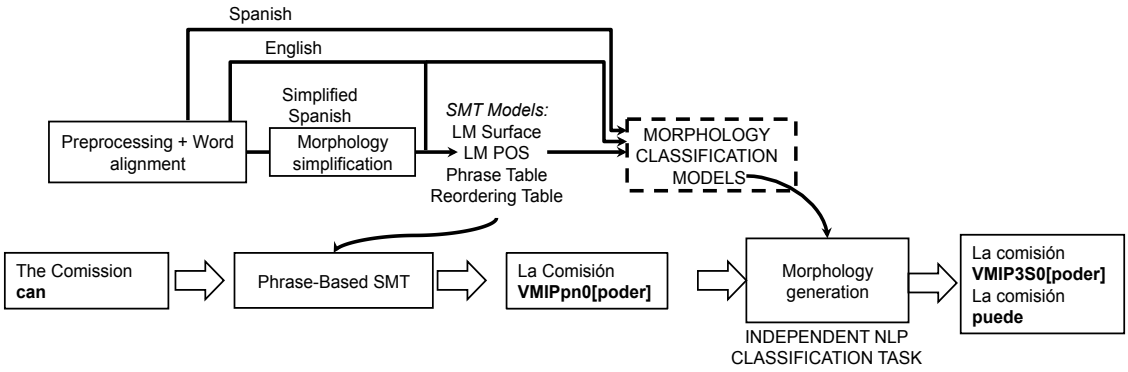

Figure 1: Flow diagram of the training of simplified morphology translation models.

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.

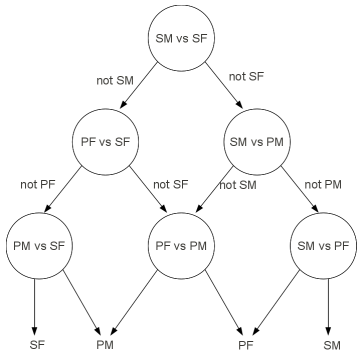

Figure 2: Decision DAG to find the best class out of four classes related to gender and number

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.

Monolingual Data Optimisation for Bootstrapping SMT Engines

Jie Jiang,[†] Andy Way,[†] Nelson Ng,[‡] Rejwanul Haque,[†] Mike Dillinger,[‡] and Jun Lu[‡]
[†]Applied Language Solutions, Delph, OL3 5FZ, UK
[‡]eBay Inc.
{jie.jiang,andy.way}@appliedlanguage.com, nng@ebay.com,
rejwanul.haque@appliedlanguage.com, {micdillinger,jlu}@ebay.com

Abstract

Content localisation via machine translation (MT) is a sine qua non, especially for international online business. While most applications utilise rule-based solutions due to the lack of suitable in-domain parallel corpora for statistical MT (SMT) training, in this paper we investigate the possibility of applying SMT where huge amounts of monolingual content only are available. We describe a case study where an analysis of a very large amount of monolingual online trading data from eBay is conducted by ALS with a view to reducing this corpus to the most representative sample in order to ensure the widest possible coverage of the total data set. Furthermore, minimal yet optimal sets of sentences/words/terms are selected for generation of initial translation units for future SMT system-building.

1 Introduction

On many occasions clients approach machine translation (MT) providers knowing that they do not have parallel data that can be used 'as is' to train statistical MT (SMT) engines. The typical approach in such circumstances is to mine the Web for data that is as 'close as possible' to the specific use-case of the client that can be used – after cleaning – as parallel training data (cf. Pecina et al. (2011)).

However, not all clients are the same, and solutions such as the above may not be appropriate in all cases. Sometimes clients have hundreds of millions of sentences of monolingual data only, which is already publicly available. Searching on the Web for 'similar' data would be meaningless, as one would typically pick up that self-same data in a web crawl.

In this paper, we describe the engagement of the language technology (LT) group at Applied Language Solutions (ALS) by eBay, the world's largest multinational online trading company, to research the feasibility of providing multilingual content from their monolingual data using MT technologies in order to facilitate a multilingual cross-border trading solution for eBay. ALS were provided with a large sample of eBay's English data – mostly user-generated content – with a view to recommending which parts of that data set were most representative of the data as a whole, and which could then be set aside for human translation so as to seed an initial parallel data set for SMT engine-building.

With the vast amount of data provided by eBay, it was essential to analyse the content prior to any further text processing, given the strong probability of a large variety of domains and text genres in the data set as a whole. We describe the corpus in detail in Section 2, but essentially the data comprised 34 separate eBay categories, with each separate data item in each category consisting of 7 fields. We needed to discover the 'closeness' of these categories, to try to reduce the number of MT engines that we would recommend be built for eBay. Once we had established this, we were required to select the minimum number of monolingual sentences to create Translation Units (TUs), which once translated (either (i) completely by hand, or (ii) via MT seeded with some manually translated data, followed by post-editing) could be used to train the engines to translate the remainder of the data, together with any new incoming

Tag Name	Total No. Sentences	No. Unique Sentences	Sentence Duplicates	No. Words	Vocabulary Size
Item Title	1,016,364	1,001,169	1.5%	8,960,683	87,580
Item Subtitle	113,007	24,040	78.73%	772,599	8,042
Item Description	32,193,213	6,194,681	80.76%	437,020,245	277,252
Payment Instructions	1,296,336	142,725	88.99%	17,663,422	14,520
Refund As	741,956	6	100%	1,399,393	13
Return Within	741,956	9	100%	1,483,912	9
Refund Details	1,447,691	108,778	92.49%	24,131,176	12,019
Totals:	37,550,523	7,471,408		491,431,430	399,435

Table 1: Monolingual Data Corpus Statistics

text. In brief, we performed monolingual data analysis on vocabulary overlap, and monolingual data clustering to find the optimal yet smallest number of domains. Finally, we extracted terminology for each cluster and then used a term-based selection process to select the optimal TUs for translation so as to provide SMT training data.

In the following sections, we introduce each of these tasks and procedures that we taken to solve the related questions. While all the experiments were carried out on English data, the methods used are generic to any language, so it is easily extensible to other eBay source-language material.

The rest of the paper is organised as follows. Section 2 describes the half a billion running words of mostly user-generated eBay content (in English) that the LT group in ALS had to deal with, with analysis performed on several levels. Section 3 describes the monolingual data clustering methods used by ALS to come up with the ideal source-side of a parallel corpus, source-language lexical and terminological data that post-translation would be optimal for SMT training that we outline in Section 4. We conclude in Section 5, and provide ways in which this analysis may be extended, in terms of other eBay user-generated data, and/or for languages other than English, or by using other techniques.

2 Monolingual Data Analysis

The monolingual data provided by eBay has 34 categories, namely: Antiques, Art, Baby, Books, Business & Industrial, Cameras & Photo, Cell Phones & PDAs, Clothing Shoes & Accessories, Coins & Paper Money, Collectibles, Computers & Networking, Crafts, Dolls & Bears, DVDs & Movies, Electronics, Entertainment Memorabilia, Everything Else, Gift Cards & Coupons, Health & Beauty, Home & Garden, Jewellery & Watches, Music, Musical Instruments, Pet Supplies, Pottery & Glass, Real Estate, Speciality Services, Sporting Goods, Sports Memorabilia Cards & Fan Shop, Stamps, Tickets, Toys & Hobbies, Travel and Video Games.

Each category contains different amounts of items. Each item contains 7 different fields (labelled with different tags), namely Item Title, Item Subtitle, Item Description, Payment Instructions, Refund As, Refund Within and Refund Details. Note that the vast majority of the data comprises user-generated content, although the three 'Refund' fields contain a lot of boilerplate material, as we will demonstrate later.

In the following sections, we firstly describe the data pre-processing performed, followed by our detailed analysis of the monolingual data, in order to compute the closeness of the different eBay categories, and the different fields within each item.

2.1 Data pre-processing

The total size of the original English data was 34.8GB in XML format. The following steps were performed to pre-process the data:

- Extraction of pure text content by stripping tags and Javascript. The text of all items was labelled with the corresponding category and tag information for further processing downstream.
- Filtering of non-English material using English dictionaries, since some of the data con-

Tag Name	Vocabulary Size before Noise Reduction	Vocabulary Size after Noise Reduction
Item Title	87,580	38,352
Item Subtitle	8,042	6,885
Item Description	277,252	71,820
Payment Instructions	14,520	9,518
Refund As	13	13
Return Within	9	9
Refund Details	12,019	8,292
Totals:	399,435	134,889

Table 2: Monolingual Data Corpus Statistics after Noise Reduction

tained characters in German, Chinese, Bulgarian, Ukrainian, Russian, French, Arabic, Sanskrit, Spanish, Greek, Armenian, French, Polish and Japanese.

- Segmentation of sentences via a set of regular expressions, and then typical MT corpus pre-processing including tokenisation and lowercasing with the Moses toolkit (Koehn et al., 2007).

2.2 Corpus statistics

After pre-processing, the statistics on the eBay data set are provided in Table 1. It is easy to see that we are dealing with huge data sizes here. From the initial 34.8GB of data, after cleaning, we see there are 37.5M sentences, the vast majority (85.7%) included in Item Descriptions. With MT as the end-task in mind, it is clear that this will only be practical if smaller samples of this overall data set can be used as training data for the language-pairs of interest to eBay. Things become much more realistic when we see in the next column that the number of unique sentences is around 7.5M; that is, 80.2% of the data is found more than once in the overall data set. It would have been reasonable to assume that the Titles, Subtitles and Descriptions are particular to the Items themselves, while Payment Instructions, Refund and Return Details are similar for each item. However, as we can see, even Item Subtitles (78.7%) and Descriptions (80.8%) contain a huge number of duplicate sentences. Only the Item Titles themselves seem to be unique.

Even so, training MT engines with 7.5M sentence-pairs is non-trivial, but of course here the data we have is only monolingual. Accordingly, for MT deployment in eBay to be practical, samples from this smaller data set still need to be selected. As we move to words, the figures become even more staggering, with a total of almost half a billion words in the set. However, only around 400K (or 0.0008%) of these words are unique. Many of these are real names, places, etc., many of which will not need to be translated, with quite a few typos contained therein to boot, which can be cleaned up prior to further processing. Accordingly, we automatically removed all typos and entries not contained in two English dictionaries and recalculated the statistics in Table 1, leading to the new numbers for 'Vocabulary Size' in Table 2.

As can be seen, the removal of the different types of noise significantly reduces the amount of vocabulary items that need to be handled. Overall, numbers decrease by 264,546, or 66%. As expected, the biggest savings are to be had for Item Descriptions (205,432 words, or 74%), but signifant reductions are seen for all other major categories: Item Titles (49228 words, or 56%), Item Subtitles (1157 words, or 14%), Payment Instructions (5002 words, or 34%), and Refund Details (3727 words, or 31%).

All of the above gives us cause for optimism, as it is clear that for multinational multilingual companies such as eBay, automation is the only way in which data sizes of the amounts shown above can be handled.

2.3 Vocabulary overlap

To investigate the closeness between each of the 34 eBay categories, we calculated the vocabulary over-

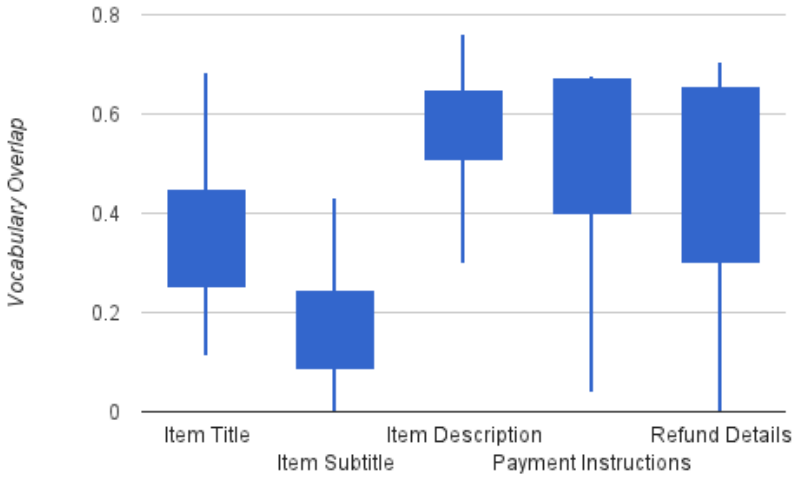

Figure 1: Vocabulary Overlap per tag between different eBay categories

lap across all categories, where the denominator is the super-set of both dictionaries. Clearly such a grid contains far too many entries to be inserted in this paper, so instead we illustrate only the maximum, minimum and standard derivations between all one-to-one vocabulary overlaps between each of the categories. The results are shown in Figure 1, and separated by different tag names. Note that we ignored Refund As and Returned Within since they are trivial to analyse with such a tiny vocabulary size.

The general observations are as follows:

- *Item Titles*: in general, vocabulary overlap is on the low side here, varying from around 11% (Real Estate ∩ Cell Phones & PDAs) to 64% (Entertainment Memorabilia ∩ Music). Some of this may be due to data sparseness, but more generally may be attributed to the free-form input in Item Titles. Some categories tend to show that they have low overlap with most other categories (e.g. Real Estate, Tickets), while others show relatively high overlaps across the board (e.g. Collectibles, Crafts, Toys & Hobbies). Generally speaking, average overlaps seem to be around 35% for this sub-part of the data.

- *Item Subtitles*: as for Titles, vocabulary overlap for Item Subtitles is low, on average around 1520%. The range of overlap varies from less than 1% (Entertainment Memorabilia ∩ Real Estate) to 43% (Electronics ∩ Cameras & Photo). Some of these very low overlaps are due to data sparseness, as this field appears to be optional, so quite low numbers of instances can be seen in the data. Overlaps are low across the board for some categories (Real Estate, Dolls & Bears, Gift Cards & Coupons, Specialty Services, Sports Memorabilia, Stamps), while for others they are relatively high, on average (Cameras & Photo, Computers & Networking, Music, Toys & Hobbies).

- *Item Descriptions*: here, for the sub-part of the Items that comprises by far the largest amount of the data, vocabulary overlap varies from around 35% (Tickets ∩ Books) to 73% (Cameras & Photo ∩ Computers & Networking). More generally overlaps of around 60% are seen. Some categories tend to show that they have low overlap with most other categories (e.g. Tickets, Real Estate), while others show relatively high overlaps across the board (such as Musical Instruments, Home & Garden, Collectibles and Cameras & Photo).

- *Payment Instructions*: The percentage of vocabulary overlap for this sub-part of the data is in general very high, with average overlaps of over 60%. These range from less than 4% (Real Estate ∩ Pottery) to 68% (Computers & Net-

working ∩ Cameras & Photo). Again, low average scores can be seen for certain categories (Real Estate, Tickets), while others have higher than average vocabulary overlaps (Toys & Hobbies, Computers & Networking).

- *Refund Details*: Again, the average vocabulary overlaps for this sub-part of the data are reasonably high, around 50% overall. Values range from 0% (Real Estate ∩ Any other category) to 71% (Cameras & Photo ∩ Electronics). As before, scores which are on the low side are seen for certain categories (Specialty Services, Gift Cards & Coupons), with others having high average overlaps (Cameras & Photo, Computers & Networking, Electronics).

In sum, vocabulary overlaps for Item Titles and Item Subtitles are on the low side, whereas high average overlaps are seen for Item Descriptions, Payment Instructions and Refund Details. Given the large overlaps at sentential level for the latter two sub-parts, these will need to be translated only once to ensure that the vast majority of future cases in these data fields will be covered and accurately translated. Where Item Titles and Item Subtitles are concerned, these will largely be covered by accurate translation of the termbanks extracted from monolingual data.

Since Item Descriptions comprise by far the largest portion of the data, the vocabulary overlap seen there is encouraging when we consider the training data samples extracted in the next section. Thus we will focus on Item Descriptions where clustering and data selection is concerned.

3 Monolingual Data Clustering

The aim in this section is to find an optimal number of MT engines to translate the monolingual data once eBay's multilingual cross-border trading solution goes live. The obvious solutions are either to use one single generic engine or 34 domain-specific engines, but these are unlikely to be the best ways forward for eBay. Accordingly, we employ data clustering techniques to identify optimal clusters based on the 34 categories for SMT engine-building.

3.1 Clustering features and algorithms

We used three different features to perform clustering on the eBay monolingual data, namely:

1. TF-IDF (Spärck Jones, 1972; Salton and McGill, 1983; Salton et al., 1983; Salton and Buckley, 1988; Wu et al., 2008),

2. Language Model (LM) Perplexity (Ponte and Croft, 1998; Song and Croft, 1999; Lv and Zhai, 2009; Manning et al., 2009; Büttcher et al., 2010),

3. Dice Coefficient (Dice, 1945; van Rijsbergen, 1979; Kondrak et al., 2003).

The problem we are confronted with here is an instance of unsupervised learning, where the data is essentially unstructured, with no annotations to guide the machine-learning process. In our case, the chosen algorithm is provided with the data alone, and has to learn some basic characteristics of that data via distributional patterns, and in this section, clustering.

We employed the following two approaches for clustering:

1. Hierarchical clustering (Press et al., 2007a; Hastie et al., 2009),

2. K-Means clustering (Kanungo et al., 2002; Press et al., 2007b).

We tried all combinations of features and algorithms, and our general findings were as follows:

- Different features tend to produce different clusters,

- Different algorithms tend to produce similar clusters. However, K-Means with a random start point tend to produce clusters with small differences.

Since our aim was to find optimal clusters for MT engine building, we chose the LM perplexity feature instead of the other two, because it is closely related to SMT performance: the lower the LM perplexity, the better the MT engine's performance with respect to translation quality, as has been widely reported (e.g. (Eck et al., 2004; Foster & Kuhn, 2007)). We also chose hierarchical clustering as we want the clustering results to be stable as opposed to changing over time.

Specifically, LM perplexity is calculated by dividing the data in each of the categories into independent training and testing sections, with random sampling from corresponding items. Language models are built on the training data, and then both within- and cross-category LM perplexities are calculated on the test data. All the perplexity scores are normalized on the in-domain perplexity score, and these are used for the distance measure of the clustering procedure in Section 3.3. We do this so that the scores are comparable across categories; this needs to be done given the varying amounts of data in each of the eBay categories. Lest there be any misunderstanding, this normalization has nothing to do with the calculation of the perplexity scores *per se*.

3.2 Optimal number of clusters

For both Hierarchical clustering and K-Means clustering, we need to determine the optimal number of clusters. For the LM Perplexity feature, average in-cluster LM Perplexity (PPL_{avg}) is used, as in formula (1):

$$PPL_{avg} = \frac{\sum_i \frac{\sum_{j,k} d(c_j,c_k)}{M_i}}{N} \quad (1)$$

where N is the number of clusters, i is the index for cluster i, M_i is the number of categories in cluster i; j and k are the indexes for categories in cluster i,and c_j and c_k are categories j and k in cluster i, $d(c_j, c_k)$ are the perplexity scores calculated by c_j text with LM model built via random sampling on c_k.

To balance the number of MT systems required and LM perplexities, the dynamics of this objective function indicate the benefits of overall LM perplexities by increasing/decreasing the total number of clusters. Accordingly, we select the number of clusters which has the biggest drop in value of the objective function.

3.3 Clustering results

The hierarchical clustering results using the LM perplexity feature on Item Description data is shown in Figure 2. Note that different clusters can be obtained with varying distance thresholds. Therefore, PPL_{avg} in formula (1) is used to determine the optimal number of clusters. PPL_{avg} and its dynamics

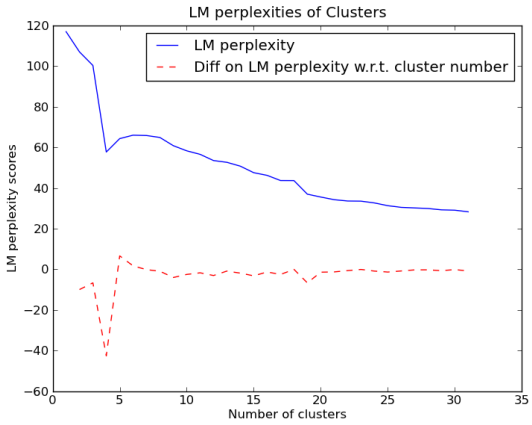

Figure 3: PPL_{avg} of different number of clusters in Hierarchical Clustering

with respect to the change in number of clusters are illustrated in Figure 3. The blue curve shows the average in-cluster LM perplexities (PPL_{avg}), and the red curve depicts the different changes in cluster numbers ($PPL_{avg}(N-1) - PPL_{avg}(N)$ at point N).

Figure 3 clearly shows that 4 clusters appears to be the best trade-off between average LM perplexity and number of clusters, as there is a big drop in PPL_{avg} from three clusters to four, and results do not change dramatically thereafter.

Accordingly, the optimal 4 clusters which result from hierarchical clustering for Item Descriptions are as follows:

- Cluster 1: "Baby", "Business Industrial", "Cameras Photo", "Cell Phones PDAs", "Computers Networking", "Dolls Bears", "Electronics", "Health Beauty", "Home Garden", "Musical Instruments", "Pet Supplies", "Sporting Goods", "Sports Mem Cards Fan Shop", "Tickets", "Toys Hobbies", "Travel", "Video Games"

- Cluster 2: "Antiques", "Art", "Books", "Clothing Shoes Accessories", "Coins Paper Money", "Collectibles", "Crafts", "DVDs Movies", "Entertainment Memorabilia", "Everything Else", "Jewelry Watches", "Music", "Pottery Glass", "Specialty Services", "Stamps"

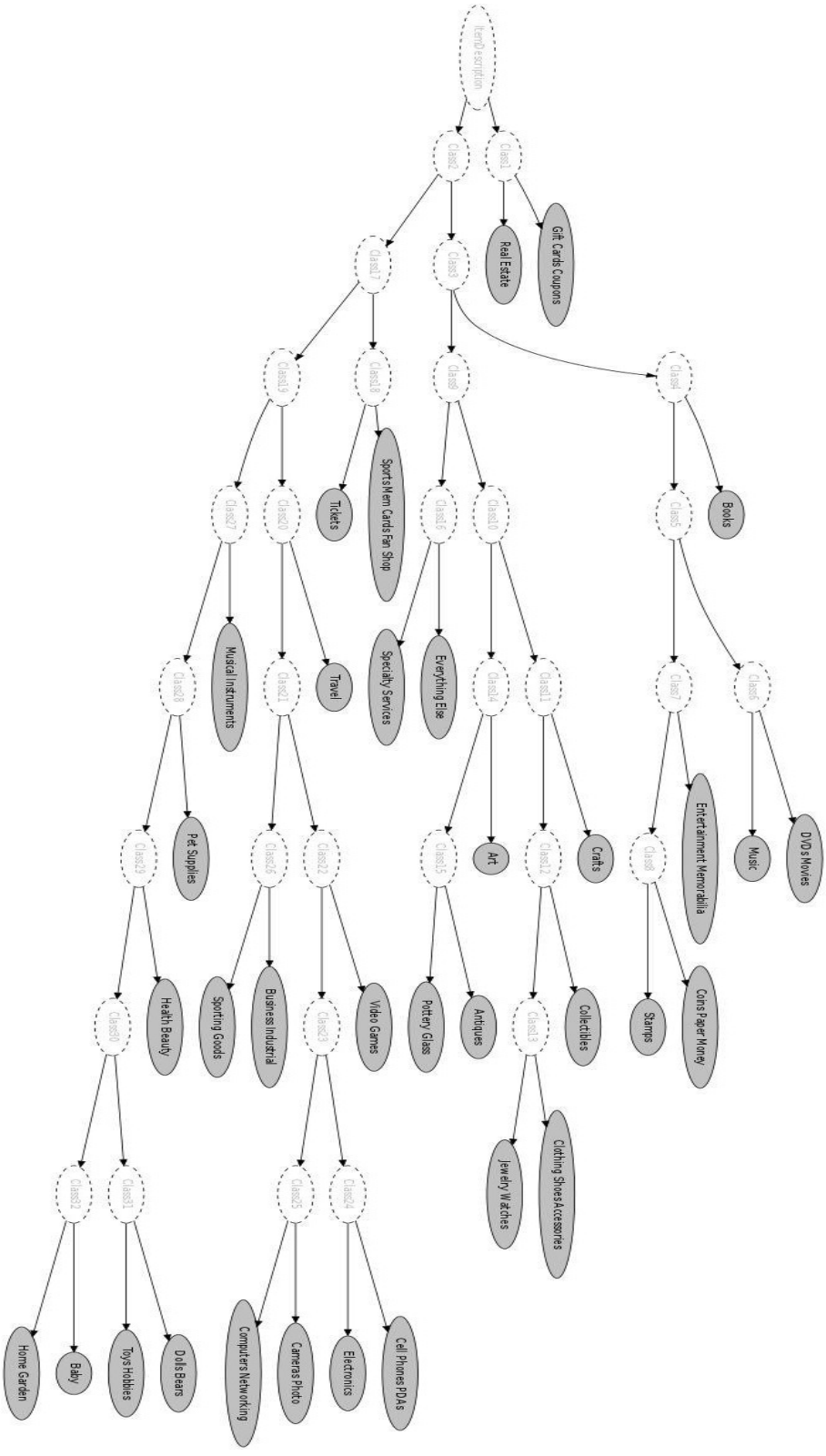

Figure 2: Tree structure of categories based on Hierarchical Clustering

- Cluster 3: “Gift Cards Coupons”
- Cluster 4: “Real Estate”

Note that two quite specific categories ‘Gift Cards & Coupons’ and ‘Real Estate’ end up in separate clusters of their own. Clearly the text content in these two categories is quite different from other categories – as we saw earlier in Section 2.3 – and should be treated differently with specific MT engines, while all other categories can be handled with two MT engines with much wider coverage in terms of domains.

With this 4-cluster information, we are able to select optimal monolingual sentence sets to seed MT corpus building, which we describe next.

4 Optimal Data for SMT Training

The results in the previous two sections are applied to find the optimal data collections for consideration as the source-side of SMT training data. As in the previous sections, we focus primarily on the selection of training data for SMT engines for Item Description data.

The first step is to extract terminology sets for each of the categories for Item Description data. TF-IDF is used to select those terms which have higher values than a given threshold. We select the threshold heuristically based on the score distribution, specifically choosing the point at which there is a significant drop in TF-IDF scores. To be more precise, we cut off when the number of terms gained at the current point is no more than X% of the previously accumulated terms. X is different across categories, and typically it varies from around 10–20%.

Then a term-based TU selection process is carried out independently for each cluster. During the selection process, we plotted the relationship between the selected words and term/vocabulary coverage. We then use these statistics to determine the optimal selection point.

We plot the relationship in a graphical representation as follows:

1. Term coverage: percentage of terms covered,
2. Vocabulary coverage: percentage of vocabulary entries covered,

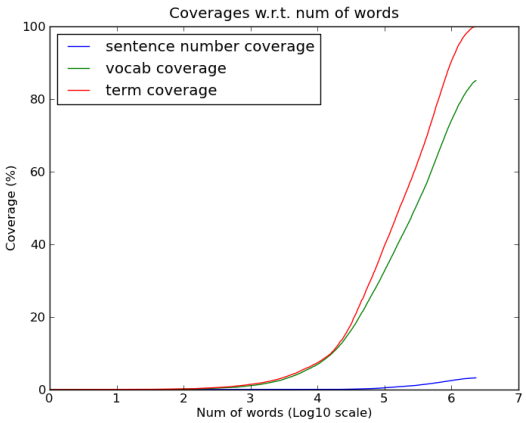

Figure 4: Data Selection Statistics for Cluster 1

3. Sentence number coverage: percentage of sentences covered.

Based on the statistics, we selected the optimal TUs where an increase in the number of words does not significantly benefit the term coverage. We use a heuristic threshold of 0.5 (on the log-scale) as the minimum increase to select the cut-off point to determine the number of words to be selected.

The selection graph of Cluster 1 is shown in Figure 4; the other three clusters have similar graphs, and are thus omitted for reasons of space. Actually the selection process is guided by the red curve in the Figure, which indicates the percentage of terms covered by the selected TUs. The optimal threshold is where term coverage is 91%.

By applying the heuristic threshold for all four clusters, we finally selected 4 different sets of TUs for the source-side of SMT training data, collected together in Table 3.

As Table 3 illustrates, across all 4 clusters for Item Descriptions, the optimal configuration only requires 2.17M words, which is less than 0.5% of the total initial set of 437 million words in Table 1.[1] Of these 2.17M words, 91% of the terms and 69% of the vocabulary are already covered by that data, showing that to obtain full terminology and vocabulary coverage, just under 4K more terms and 49K extra vocabulary items need to be translated in addition to

[1]This estimate is on the optimistic side, as it assumes that the target language has a comparable vocabulary size to English, which is not true for many (or most) languages.

No. Words	Term Coverage	No. Terms Uncovered	Vocab Coverage	No. Vocab Uncovered	Total Words
2,176,009	91%	3,959	68.7%	49,022	2,228,990

Table 3: Summary of Optimal TU selection

the source-side sentential data. This reduction in TU selection massively reduces the amount of data to be considered during the MT system-building process.

5 Conclusions and Further Work

In this paper, the LT group from ALS, a well-known translation service provider, analysed a huge sample of monolingual content sampled from eBay, the largest multinational online trading company. For 34 different categories with 7 different fields per Item, we calculated the sentence duplication and vocabulary overlaps to show that for most of the data, MT can be a suitable solution for content localisation where training data does not exist *a priori*. We then obtained the optimal number of clusters via monolingual data clustering, and then selected optimal sentences/words/terms that can be used to create an initial parallel corpus to train SMT engines for each cluster. Compared with the original huge data size, we demonstrated that only a very small amount of the total words need to be translated in the optimal training corpus to ensure maximum coverage.

What is absolutely certain is that for multinational companies like eBay who are interested in a multilingual solution, automation is key; handling data sizes of the magnitude that ALS had to deal with in this paper would be unthinkable by humans alone.

As far as extensions to this work are concerned, similar analysis could be performed for source languages other than English, since all of the techniques applied to the eBay English data can be easily extended to other languages with only small modifications. Meanwhile, further eBay user-generated content, such as buyer-seller interactions, and review guides and catalogue information are interesting data for further analysis.

In addition, we would like to see how cross-lingual information retrieval techniques (e.g. (Snover et al., 2008)) to automatically select parallel data compare to our method. Furthermore, given the similarity of the use-case described here to patent translation, approaches to domain-adaptation (e.g. (Banerjee et al., 2011)) or multi-task learning for patent translation (e.g. (Tinsley et al., 2010; Ceausu et al., 2011; Wäschle & Riezler, 2012) might be applicable to the e-commerce domain.

We are also interested in deeper levels of analysis arising from the study carried out in this paper. For example, we could have used the three clustering metrics together to seek further corroboration of the clusters that we ended up with, or used other techniques altogether (cf. (Mandal et al., 2008; Biçici & Yuret, 2011)). Finally, of course, the next stage is to build, apply, and evaluate the SMT engines constructed on the basis of the recommendations provided here on new, unseen eBay data, and compare the results of other method against these other possible approaches.

Acknowledgments

Many thanks to the anonymous reviewers, especially for the suggestions regarding related work.

References

Banerjee, P., S. Naskar, J. Roturier, A. Way and Josef van Genabith. 2011. Domain Adaptation in Statistical Machine Translation of User-Forum Data using Component Level Mixture Modelling. In *Proceedings of Machine Translation Summit XIII*, Xiamen, China, pp. 285–292.

Biçici, E. and D. Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT '11*, Edinburgh, Scotland, pp. 272–283.

Büttcher, S., C. Clarke and G. Cormack. 2010. *Information Retrieval: Implementing and Evaluating Search Engines*. Cambridge MA: MIT Press, pp. 289–291.

Ceausu, A., J. Tinsley, J. Zhang and A. Way. 2011. Experiments on domain adaptation for patent machine translation in the PLuTO project. In *Proceedings of the 15th Annual Meeting of the European Association for Machine Translation*, Leuven, Belgium, pp. 21–28.

Dice, L. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26(3): 297–302.

Eck, M., S. Vogel and A. Waibel. 2004. Language Model Adaptation for Statistical Machine Translation based on Information Retrieval. In *LREC-2004: Fourth International Conference on Language Resources and Evaluation, Proceedings*, Lisbon, Portugal, pp. 327–330.

Foster, G. and R. Kuhn. 2007. Mixture-model adaptation for SMT. In *ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 128–135.

Hastie, T., R. Tibshirani and J. Friedman. 2009. *The Elements of Statistical Learning* (2nd ed.). New York: Springer, pp. 520–527.

Kanungo, T., D. Mount, N. Netanyahu, C. Piatko, R. Silverman and A. Wu. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence* **24**: 881–892.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *ACL 2007: Proceedings of Demo and Poster Sessions*, Prague, Czech Republic, pp. 177–180.

Kondrak, G., D. Marcu and K. Knight. 2003. Cognates Can Improve Statistical Translation Models. In *Proceedings of HLT-NAACL 2003: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada, pp. 46–48.

Lv, Y. and C-Z. Zhai. 2009. Positional Language Models for Information Retrieval. in *Proceedings of the 32nd International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, Boston MA., pp. 299–306.

Mandal, A., D. Vergyri, W. Wang, J. Zheng, A. Stolcke, G. Tur, D. Hakkani-Tür and N. Ayan. 2008. Efficient Data Selection for Machine Translation. In *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology*, Goa, India, pp. 261–264.

Manning, C., P. Raghavan and H. Schütze. 2009. An Introduction to Information Retrieval. In C. Manning & H. Schütze (eds.) *Foundations of Statistical Natural Language Processing*, Cambridge University Press, pp. 237–240.

Pecina, P., A. Toral, A. Way, P. Prokopidis and V. Papavassiliou. 2011. Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 15th Annual Meeting of the European Association for Machine Translation*, Leuven, Belgium, pp. 297–304.

Ponte, J. and B. Croft. 1998. A Language Modeling Approach to Information Retrieval. In *SIGIR '98, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 275–281.

Press, W., S. Teukolsky, W. Vetterling and B. Flannery. 2007a. Section 16.4. Hierarchical Clustering by Phylogenetic Trees. *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). Cambridge: Cambridge University Press.

van Rijsbergen, K. 1979. *Information Retrieval*. London: Butterworths.

Salton, G. and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* **24**(5): 513–523.

Salton, G., E. Fox and H. Wu. 1983. Extended Boolean Information Retrieval. *Communications of the ACM* **26**(11): 1022–1036.

Salton, G. and M. McGill. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Song, F. and B. Croft. 1999. A General Language Model for Information Retrieval. In *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*. Berkeley CA., pp. 279–280.

Snover, M., B. Dorr and R. Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *EMNLP 2008: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, HI, USA, pp. 857–866.

Spärck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* **28**(1): 11–21.

Tinsley, J., A. Way and P. Sheridan. 2010. PLuTO: MT for Online Patent Translation. In *AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas, Proceedings*, Denver, CO., pp. 435–442.

Wäschle, K. and S. Riezler. 2012. Structural and topical dimensions in multi-task patent translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL-12*, Avignon, France, pp. 818–828.

Wu, H., R. Luk, K. Wong and K. Kwok 2008. Interpreting TF–IDF term weights as making relevance decisions. *ACM Transactions on Information Systems* **26**(3): 1–37.

Shallow and Deep Paraphrasing for Improved Machine Translation Parameter Optimization

Dennis N. Mehay and Michael White
The Ohio State University
Columbus, Ohio, USA
{mehay,mwhite}@ling.ohio-state.edu

Abstract

String comparison methods such as BLEU (Papineni et al., 2002) are the de facto standard in MT evaluation (MTE) and in MT system parameter tuning (Och, 2003). It is difficult for these metrics to recognize legitimate lexical and grammatical paraphrases, which is important for MT system tuning (Madnani, 2010). We present two methods to address this: a shallow lexical substitution technique and a grammar-driven paraphrasing technique. Grammatically precise paraphrasing is novel in the context of MTE, and demonstrating its usefulness is a key contribution of this paper. We use these techniques to paraphrase a single reference, which, when used for parameter tuning, leads to superior translation performance over baselines that use only human-authored references.

1 Introduction

Because of their speed, simplicity and portability, string comparison methods such as BLEU (Papineni et al., 2002) have become the de facto standard in automatic MTE, as well as in parameter tuning regimes that make heavy use of evaluation, such as MERT (Och, 2003). Although BLEU can be effective for measuring system-level differences among similar systems (Papineni et al., 2002), the surface nature of BLEU's string comparisons makes it difficult to recognize legitimate morphological, syntactic, lexical and paraphrase variation (Callison-Burch et al., 2006), and such recognition is important for MT tuning (Madnani, 2010). One way to address this issue is to to devise extensions to BLEU (Zhou et al., 2006) or to develop new string-based metrics that account for paraphrase and lexical synonymy such as Meteor (Denkowski and Lavie, 2011). Another way is to pad the reference set automatically with paraphrases of the original references (Owczarzak et al., 2006), possibly also doing so in a way that targets each hypothesis under evaluation (Madnani, 2010). The approach described here draws inspiration from both of these tactics, and uses the Meteor metric and a large corpus of n-grams to extend a reference set in a way that is targeted to the output of a baseline system (Section 2). In addition, we generate word-order variants of both the original and the lexically paraphrased references by using a high-precision, grammar-driven parsing and realization system (Section 3). The use of sentence-level paraphrase (Madnani, 2010) — or a rough-and-ready approximations to it (Dyer et al., 2011) — is not new to MTE or parameter tuning. Using deep, grammatically-driven paraphrase, however, is novel in the context of MTE, and demonstrating its usefulness for parameter tuning is a key contribution of this paper.

Targeted lexical substitutions produce reference translations that are more likely **reachable** and **focused** (relevant) w.r.t. a particular translation system being tuned, while grammatical paraphrase helps ensure **correctness**. These are three qualities that Madnani (2010) has argued are important for MT parameter tuning. In a MERT tuning scenario, we find that both paraphrase methods (lexical and grammatical) lead to improved translation results on two held-out validation sets.

2 A Simple Method for Targeted Lexical Paraphrase

Arguably, metrics such as Meteor, which have high correlation to human judgments (Owczarzak, 2008; Denkowski and Lavie, 2011), should be incorporated into system building pipelines. But BLEU is often the metric of evaluation in cross-system comparisons and hence is usually optimized.[1] The technique presented here allows Meteor's lexical knowledge to be injected into the reference set, and therefore into a BLEU-based tuning regime.

Meteor (Denkowski and Lavie, 2011) aligns hypotheses with their references in a greedy multi-stage process that matches with word forms, then stems, then lexical synonyms, then automatically derived, multiword paraphrases (Bannard and Callison-Burch, 2005). This process is primarily intended to be used directly for evaluation, but it can also be used for other purposes. For example, a targeted paraphrase of one string can be created by substituting into it some of the aligned words from the other string. In this way a reference translation can be modified to target MT system outputs, and thereby extend the reference set in a way that is relevant for retuning that system, as in Figure 1.

Figure 1 depicts a correct paraphrase, but nothing is preventing ungrammatical substitutions such as "...have been **form** election alliances". To address this concern, we follow Chang and Clark (2010) and only permit substitutions that overlap with their context in a way that forms n-grams that were observed in the large corpus of n-grams (or in the original reference). As an example, picking a setting of n=2, this filter would only allow the substitution of '**the parties**' for the original phrase '**these parties**' if the edge bigrams '**all the**' and '**parties have**' occur in either the n-gram corpus or the original references. Note that not filtering with n-grams is similar to Owczarzak et al.'s (2006) method, where paraphrases were mined from the reference set.

Chang and Clark (2010) use the Google n-gram corpus (Brants and Franz, 2006), and they find that a value of n=2 performs best, with an F-measure of 76%. We follow their lead here. They also employ a second, parser-based filter in order to raise the precision of their paraphrase substitutions. We do not use this second filter, as it was not designed to address multiple substitutions in the same sentence.

[1] Och (2003) showed empirically that the metric used for tuning was the one that systems performed best on at test.

3 Precise, Grammatically-Driven, Deep Paraphrasing with OpenCCG

In addition to lexical and multiword paraphrases, string-based MTE metrics also struggle to account for grammatically licensed word-order variation. To enumerate grammatically licensed paraphrases of a reference, we use OpenCCG, an open-source parsing and realization system.[2] OpenCCG features a symbolic-statistical chart parser and surface realizer (White and Rajkumar, 2012). The OpenCCG parser consumes strings and produces semantic dependency graphs (White, 2006), which abstract away from the order of the string. The realizer consumes these graphs and enumerates string realizations that cover all nodes in the graph subject to the grammatical constraints of the CCG syntax. When the parser and realizer are chained end-to-end, OpenCCG becomes a precise, grammatically-driven paraphrase system. As an example, consider the reference *The minister did not however name any associated agency*, which OpenCCG paraphrases as *The minister, however, did not name any associated agency*. This provides a better (uncased) match to one system's output during tuning: *the minister said , however , did not name any assistant organization*. We apply this method both to the original references and to those that have been paraphrased by the method in Section 2.

4 Improved Parameter Optimization in Phrase-Based MT

We use both paraphrase methods described above to test the effects of paraphrasing on MT tuning performance in an Urdu-English translation task. We train phrase-based systems using the Moses toolkit (Koehn et al., 2007). All systems use "hierarchical" lexicalized reordering (Galley and Manning, 2008) and a large distortion limit of 15 to account for the differences in Urdu and English word order. We tune system parameters using MERT with BLEU as the tuning metric. For each experimental condition, we run MERT three times and test for significance

[2] http://openccg.sourceforge.net

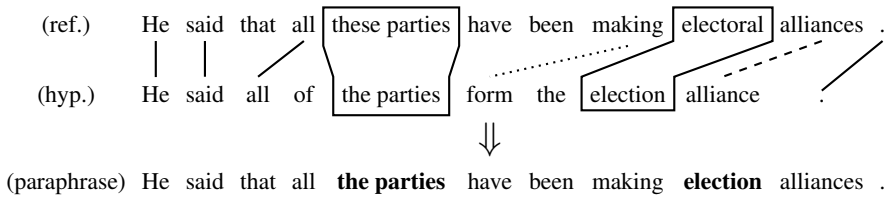

Figure 1: **Top:** Meteor's (v. 1.3) multi-stage alignment using exact matches (solid line), Porter stemming (dashed line), WordNet synonymy (dotted line) and paraphrases (solid outline). **Bottom:** A potential paraphrase.

using Clark et al.'s (2011) method, which helps to control for MERT's inherent instability. Our translation data is the OpenMT Urdu-English training set, which was chosen because its corresponding evaluation reference sets have four references per source segment, allowing a direct comparison with multiple human-authored references. For development tuning and testing, we split the OpenMT-08 evaluation data [LDC2010T21] into two balanced halves of ≈900 segments each. For further test data, we use the whole of the OpenMT-09 evaluation data [LDC2010T23], which has 1,792 segments. Results are computed on the evaluation sets using all four original, human-authored references (i.e., with no paraphrase). All that varies between conditions is the reference sets that are used for tuning.

We tune using the following conditions:[3]

BASELINE-4: 4 human-authored references.

BASELINE-1: 1 "BLEU-best", human-authored reference (see below).

PARASUBS2G: BASELINE-1 + substitutions via the method in Section 2 (with bigram filter).

PARASUBSNOFILT: Like PARASUBS2G, but **without** the bigram filter.

REVR: BASELINE-1 + 2-best reverse realizations of it (distinct from the original) via the method from Section 3.

PARASUBS2GREVR: BASELINE-1 + paraphrase substitutions (with bigram filter) + 2-best reverse realizations thereof.

PARASUBSNOFILTREVR: Like PARASUBS2G-REVR, but **without** the bigram filter.

[3]Note that it may be possible to obtain further gains by using additional n-best realizations.

The second, "BLEU-best" condition is obtained by selecting, for each segment in the development tuning set, the single reference that has the highest BLEU score w.r.t. the other references that are distinct from it. This approximates picking the reference that is "best" w.r.t. the other references without rewarding exact duplicates. Table 1 lists tuned system results. As expected, adding lexical paraphrases using the method from Section 2 improves both BLEU and Meteor performance. What was unexpected is that **not** applying the bigram filter leads to higher scores than applying it does. This may be because, in addition to filtering out incorrect lexical substitutions, the bigram filter also blocks *correct* substitutions, when their edge bigrams are not found in the Google n-grams. Also unexpected was that the "BLEU-best" single reference case and the cases where it was paraphrased were superior to the multi-reference condition, in contrast to what Madnani (2010) found in a four reference scenario for Chinese-English translation. This might be due to our method of choosing a single reference, or to a peculiarity of the Urdu-English data set. Nevertheless, the combination of Meteor-driven lexical substitution and OpenCCG parsing and realization achieved high Meteor scores and the highest BLEU scores in all cases. Because BLEU has been found to be sensitive to translation fluency (Owczarzak, 2008), we speculate that the higher BLEU scores may indicate that the grammatical paraphrase tuning method is improving the fluency of the output.

5 Conclusion

We have shown how to extend a set of reference translations using lexical and multiword paraphrase substitution, grammatically licensed reordering, or a combination of the two. All of these techniques lead

	DEVTEST		OPENMT 2009 EVAL	
	BLEU↑	Meteor↑	BLEU↑	Meteor↑
BASELINE-4	26.5	28.1	30.9	29.9
BASELINE-1	26.6	28.5*	31.1	30.3*
PARASUBS2G	26.6	28.7*	31.1	30.4*
PARASUBSNOFILT	26.8	**28.8***	31.7*	**30.7***
REVR	26.8	28.7*	31.8*	30.5*
PARASUBSNOFILTREVR	26.9*	28.7*	**32.1***	30.5*
PARASUBS2GREVR	**27.4***	**28.8***	32.0*	30.5*

Table 1: Uncased BLEU and Meteor scores on the Urdu-English validation sets. Results significantly better than **Baseline-4** ($p \leq 0.05$) have a '*', and the highest scores are boldfaced. See above for abbreviations.

to improved MT parameter tuning, as compared to using only human-authored translations. In future work, we plan to extend these results to other language pairs and to measure correlations to human judgments.

Acknowledgments

This work was supported in part by the Air Force Research Laboratory under a subcontract to FA8750-09-C-0179.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the ACL*, Ann Arbor, MI, USA.

Thorsten Brants and Alex Franz. 2006. Google Research Web 1T 5-gram Corpus Version 1.1 (LDC2002S28). Linguistic Data Consortium, Philidelphia, PA, USA.

Chris M. Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the EACL*, Trento, Italy.

Ching-Yun Chang and Stephen Clark. 2010. Linguistic steganography using automatically generated paraphrases. In *Proceedings of HLT:ACL*, Los Angeles, CA, USA.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of ACL: HLT*, Portland, OR, USA.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of WMT-11*, Edinburgh, U.K.

Chris Dyer, Kevin Gimpel, Jonathan H. Clark, and Noah A. Smith. 2011. The CMU-ARK German-English Translation System. In *Proceedings of WMT-11*, Edinburgh, U.K.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of EMNLP-08*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL, Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic.

Nitin Madnani. 2010. *The Circle of Meaning: From Translation to Paraphrasing and Back*. Ph.D. thesis, University of Maryland, College Park, MD, USA.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the ACL*, Sapporo, Japan.

Karolina Owczarzak, Declan Groves, Josef Van Genabith, and Andy Way. 2006. Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. In *Proceedings of WMT-06*.

Karolina Owczarzak. 2008. *A Novel Dependency-Based Evaluation Metric for Machine Translation*. Ph.D. thesis, Dublin City University, Dublin, Ireland.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL*, Philadelphia, PA, USA.

Michael White and Rajakrishnan Rajkumar. 2012. Minimal dependency length in realization ranking. In *Proceedings of the EMNLP-12/Computational Natural Language Learning*, Jeju Island, Korea.

Michael White. 2006. Efficient Realization of Coordinate Structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In *Proceedings of EMNLP*, Sydney, Australia.

Two stage Machine Translation System using Pattern-based MT and Phrase-based SMT

Jin'ichi Murakami
Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University, Japan
murakami@ike.tottori-u.ac.jp

Takuya Nishimura
Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University, Japan
s062044@ike.tottori-u.ac.jp

Masato Tokuhisa
Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University, Japan
tokuhisa@ike.tottori-u.ac.jp

Abstract

We have developed a two-stage machine translation (MT) system. The first stage consists of an automatically created pattern-based machine translation system (PBMT), and the second stage consists of a standard phrase-based statistical machine translation (SMT) system. We studied for the Japanese-English simple sentence task.

First, we obtained English sentences from Japanese sentences using an automatically created Japanese-English pattern-based machine translation. We call the English sentences obtained in this way as *"English"*. Second, we applied a standard SMT (Moses) to the results. This means that we translated the *"English"* sentences into English by SMT. We also conducted ABX tests(Clark, 1982) to compare the outputs by the standard SMT (Moses) with those by the proposed system for 100 sentences.

The experimental results indicated that 30 sentences output by the proposed system were evaluated as being better than those outputs by the standard SMT system, whereas 9 sentences output by the standard SMT system were thought to be better than those outputs by the proposed system. This means that our proposed system functioned effectively in the Japanese-English simple sentence task.

1 Introduction

Machine translation (MT) systems have been extensively studied, and there are now three generations of this technology. The first generation consists of rule-based MT (RBMT) systems. A pattern-based MT (PBMT) system is a kind of RBMT system. The second generation consists of example-based machine translation systems, and the third generation consists of statistical machine translation (SMT) systems, which have become very popular. Many versions of SMT systems have been introduced. An early SMT system was based on word-based models (IBM $1 \sim 5$(Brown et al., 1993)). Recent statistical MT systems have usually used phrase-based models.

However, some problems arise with phrase-based SMT. One problem is the language model. Generally, an N-gram model is used as the language model. However, this kind of model includes only local language information and does not include grammatical information. To solve this problem, we developed a two-stage MT system. The first stage consists of an automatically created PBMT system, and the second stage consists of a standard SMT system.

For Japanese-English translation, the first stage consists of Japanese-English PBMT. In this stage, we obtain *"English"* sentences from Japanese sentences. Our aim is to produce grammatically correct *"English"* sentences. However, these *"English"* sentences sometimes have low levels of fluency because they were obtained using an automatically created PBMT. In the second stage, we use a standard SMT system. This stage involves *"English"* to English machine translation. With this stage, our aim is to revise the outputs of the first stage in order to improve fluency.

We developed a PBMT system for the first stage using "train-model.perl"(Koehn et al., 2007). We

also developed a standard SMT system for the second stage using general SMT tools such as "Moses" (Koehn et al., 2007). We used these data and tools to translate Japanese-English simple sentences.

We obtained a Bilingual Evaluation Understudy (BLEU) score(Papineni et al., 2002) of 0.1821 with our proposed system. In contrast, we obtained a BLEU score of 0.2218 in the Japanese-English simple sentences using a standard SMT system (Moses). This means that the proposed system was not effective for automatic evaluation in the Japanese-English simple sentence task.

However, we conducted ABX tests(Clark, 1982) to compare the output of the standard SMT system (Moses) and the output of the proposed system for 100 sentences. The results indicated that 30 sentences of the proposed system were thought to be better than those of the standard SMT system, and 9 sentences of the standard SMT system were thought to be better than those of the proposed system. This means that our proposed system was effective in the Japanese-English simple sentence task for human evaluation.

2 Concept of Two-Stage Machine Translation

One problem with phrase-based statistical machine translation is the language model. Generally, an N-gram model is used as the language model. However, this model includes only local language information and does not include grammatical information. We studied hierarchical phrase-based statistical machine translation (HSMT) (Li et al., 2009) as a way to include grammatical information. However, HSMT analysis is similar to that of context-free grammars (CFG). We believe that such analysis complicates statistical machine translation by adding too many parameters. Therefore, it is unreliable and does not perform well, especially for the small amount of training data. On the contrary, PBMT is well known and has been extensively studied. Normally, PBMT is simple and has few parameters compared to CFG-based MT, and the output of PBMT contains grammatical information. However, there is a trade-off between the coverage of input sentences and the translation quality in the PBMT results. If we obtain good translation quality, then the coverage of RBMT for input sentences is low in the translation. If we obtain high coverage for input sentences, the translation quality is low.

We propose a two-stage MT system to overcome these problems. We developed a PBMT system for the first stage. This PBMT system had low coverage and high quality. When Japanese sentences were translated using this system, the quality of the output was good, and the outputs contained grammatical information. When not using the PBMT system to translate Japanese sentences, we used a standard SMT system. Therefore, we can obtain good quality from the entire system. Also, PBMT systems are usually created manually, which results in a huge labor cost. Therefore, we developed an automatically created PBMT system. However, this automatic PBMT output sometimes had less fluency, so we added SMT after PBMT to improve the fluency. In this system, we used PBMT in the pre-processing stage of SMT.

3 Related Work

Two-stage MT systems have been proposed before (Xu and Seneff, 2008), (Ehara, 2007), (Dugast et al., 2007), (Simar et al., 2007). L. Dugast, et al. (Dugast et al., 2007) and M. Simard, et al. (Simar et al., 2007) applied SYSTRAN and SMT for Japanese-English translation. Their concept was to use SMT as a post-process for SYSTRAN. The results of these studies indicated that these systems are more effective than using SYSTRAN or SMT alone. In M. Simard's research (Simar et al., 2007), the BLEU score was 0.2598 for SMT and 0.2880 for SYSTRAN + SMT in English-Japanese translation, and 0.2517 for SMT and 0.2679 for SYSTRAN + SMT in Japanese-English translation. Ehara(Ehara, 2007) reported on the same system for Japanese-English translation of a patent task. The BLEU score was 0.2821 for SMT and 0.2921 for RBMT + SMT. Ehara's RBMT system was a commercial Japanese-English system. For these systems, SMT was used in the post-process for RBMT, which means that SMT was used as a means of language adaptation. Also, these RBMT systems were created by hand, so they were expensive to build.

4 Pattern-Based Machine Translation

We developed an automatically created Japanese-English pattern-based machine translation system using "train-model.perl" (Koehn et al., 2007). Our system is divided into two processes. One is a process to form Japanese-English patterns, and the other is a decoding process. The details of these two processes are described below.

4.1 Japanese-English Patterns

We developed the following process for forming Japanese-English patterns.

1. Parallel Japanese-English Corpus

   We prepare Japanese-English parallel sentences for training. Example sentences are listed in Table 1.

Table 1: Parallel Japanese-English Corpus

Japanese sentence	信号 は 赤 だ.
English sentence	The light was red.

2. Japanese-English Phrase Table

   We construct a Japanese-English phrase table using train-model.perl (Koehn et al., 2007). An example Japanese-English phrase table is given in Table 2.

Table 2: Example of Japanese-English Phrase Table

Ex.1	信号 |||The light |||0.5 0.07 0.5 0.2
Ex.2	信号 は |||lights |||0.01 0.06 0.03 0.04
Ex.3	赤 |||red |||0.2 0.1 0.2 0.3
Ex.4	赤 だ |||red |||0.3 0.2 0.2 0.2

3. Japanese-English High Probability Phrase Table

   We deleted the low-probability Japanese-English phrase table (Table 2), in which the threshold was 0.1. We call the resulting table a Japanese-English high-probability phrase table (HPPT). An example of an HPPT is presented in Table 3.

Table 3: Example of Japanese-English High Probability Phrase Table

Ex.1	信号 |||The light |||0.5 0.07 0.5 0.2
Ex.3	赤 |||red |||0.2 0.1 0.2 0.3
Ex.3	赤 だ |||red |||0.3 0.2 0.2 0.2

4. Japanese-English Patterns

   We used Japanese-English parallel sentences (Table 1) and the Japanese-English HPPT (Table 3) to form Japanese-English patterns. Note that all possible Japanese-English patterns were generated. Therefore, one or more Japanese-English patterns were generated from one Japanese-English parallel sentence. Example Japanese-English patterns are listed in Table 4.

Table 4: Japanese-English Patterns

Ex.1	Japanese pattern	X1 は X2 だ .
	English pattern	X1 was X2 .
Ex.2	Japanese pattern	X1 は X2 .
	English pattern	X1 was X2 .

Figure 1 shows a flowchart for forming Japanese-English patterns.

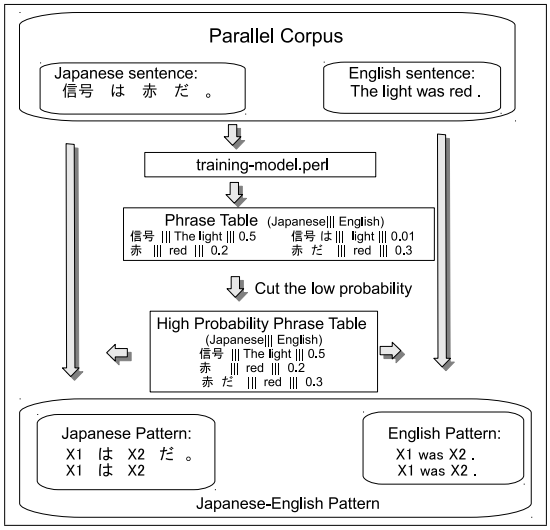

Figure 1: Formation of Japanese-English Patterns

4.2 Decoding Pattern

The decoding process is as follows.

1. Input Japanese Sentences

   We prepare input Japanese sentences. An example sentence is given in Table 5.

Table 5: Japanese Sentence

郵便 局 は どこ に 有り ます か ？

2. Search Japanese Pattern and Output English Pattern

   We search for a Japanese pattern that is matched with the input Japanese sentence using Japanese patterns and the HPPT (section 4.1). Then we obtain English patterns. Example Japanese-English patterns are listed in Table 6. Also, an example Japanese-English HPPT is shown in Table 7.

Table 6: Japanese-English Patterns

Ex.1	Japanese Pattern	X1 X2 は どこ に 有り ます か ？
	English Pattern	Where's the X2 X1 ?
Ex.2	Japanese Pattern	X2 は どこ に 有り ます か ？
	English Pattern	Where is a X2 ?

Table 7: Japanese-English High-Probability Phrase Table

Ex.1	局 |||post |||0.5 0.07 0.5 0.21
Ex.2	郵便 |||postal |||0.4 0.031 0.2 0.11
Ex.3	郵便 局 |||postal service |||0.1 0.07 0.1 0.01

3. Generate English Sentences

   We generate English sentences using the English pattern and Japanese-English High-Probability phrase tables. Note that all possible English sentences are generated. Therefore, multiple English sentences are generated from an input Japanese sentence. Example English sentences are listed in Table 8.

Table 8: Generated English Sentences

Ex.1	Where's the post office ?
Ex.2	Where is a post station ?

4. Select English Sentence.

   We select one English sentence from the multiple generated English sentences using 3-gram. We used the n-gram-count in the Stanford Research Institute Language Model (SRILM) toolkit (Stolcke, 2002) and used "-ukndiscount -interpolate" as the smoothing parameter.

   An example of an English sentence that might be selected is shown in Table 9.

Table 9: Select English Sentence

Where is a post station ?

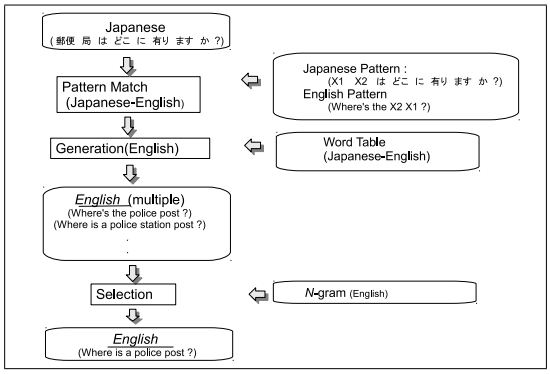

Figure 2: Decoding of Pattern-Based Machine Translation

Figure 2 shows the process of decoding English sentences for PBMT. We refer to the output of the proposed pattern-based machine translation as *"English"* sentences.

5 Overview of Proposed Machine Translation System for Training

The training model of our proposed machine translation system has three parts. The first process involves constructing an *"English"*-English phrase table, the second process involves constructing a Japanese-English phrase table, and the third part involves constructing a language model (N-gram).

5.1 *"English"*-English phrase table

"English"-English phrase tables are constructed as follows.

1. Parallel Corpus

   We prepare a Japanese-English parallel corpus.

2. Pattern-Based Machine Translation

   We use Japanese-English PBMT. Thus, we obtain *"English"* sentences from Japanese sentences. These *"English"* sentences are pairs of English sentences.

3. *"English"*-English phrase tables

   We construct *"English"*-English phrase tables using Giza++ (Och and Ney, 2003) and train-model.perl (Koehn et al., 2007) from the *"English"* sentences (outputs of Japanese-English PBMT) and English sentences (from the parallel corpus).

Figure 3 is a flow chart that shows how *"English"*-English phrase tables are constructed.

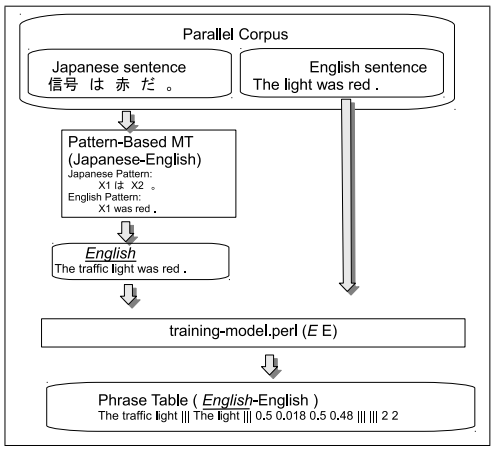

Figure 3: Flowchart for constructing *"English"*-English Phrase Tables

5.2 Japanese-English Phrase Table

We construct a Japanese-English phrase table using Giza++ (Och and Ney, 2003) and train-model.perl (Koehn et al., 2007) using the Japanese-English parallel corpus. Figure 4 shows a flow chart for constructing Japanese-English phrase tables.

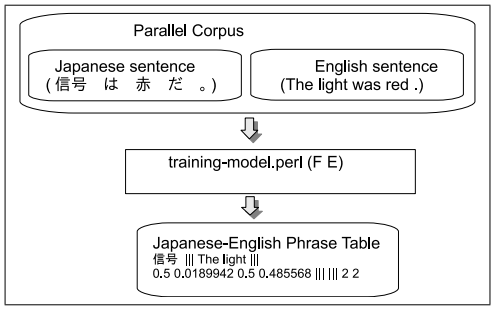

Figure 4: Flowchart for Constructing Japanese-English Phrase Tables

5.3 Language Model (N-gram).

We calculated the 5-gram model using the n-gram-count in the SRILM toolkit (Stolcke, 2002) and used "-ukndiscount -interpolate" as the smoothing parameter.

6 Overview of Proposed Machine Translation System for Decoding

The decoding process is as follows.

1. Test Corpus

   We prepare Japanese test sentences.

2. Pattern-Based Machine Translation

   We use a Japanese-English Pattern-Based Machine Translation. If an input Japanese sentence matches the Japanese patterns, we can obtain a translated *"English"* test sentence.

3. *"English"*-English Statistical Machine Translation

   Using an *"English"*-English phrase table, N-gram model, and Moses (Koehn et al., 2007), we decode the *"English"* test sentence. This involves *"English"*-English translation, resulting in an English sentence.

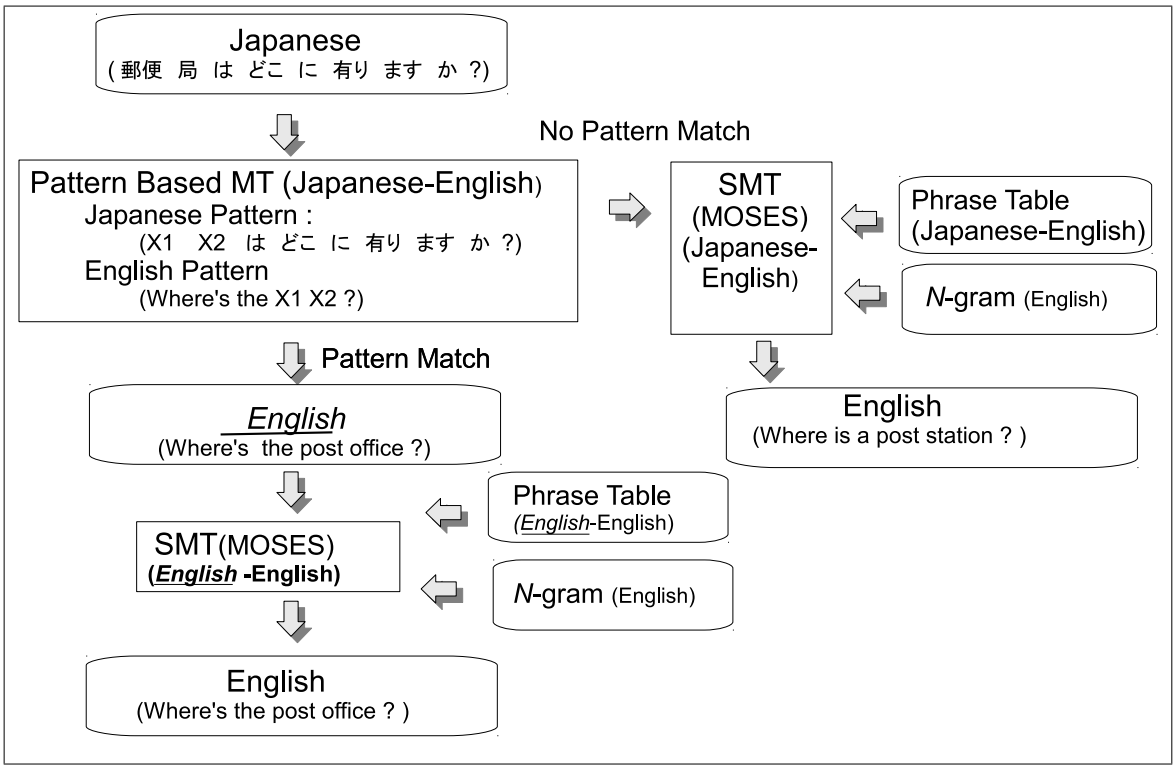

Figure 5: Flowchart of Decoding Process

4. Japanese-English Statistical Machine Translation System

   If an input Japanese sentence does not match the Japanese patterns, we conduct a standard Japanese-English SMT using a Japanese-English phrase table and N-gram model to obtain an English sentence.

Figure 5 shows a flowchart of the decoding process.

7 Experiments with our Machine Translation System

7.1 Japanese-English Simple Sentence Task

We collected a large number of Japanese-English parallel sentences from many electronic media sources. Then we selected simple sentences from these Japanese-English parallel sentences(Murakami et al., 2007). We used these Japanese-English simple sentences for training and development and test data.

Also we formed these sentences as follows. We used the English punctuation system, which means we changed "," and "." to " " and " ". And, we did not take into account English case forms. Also, we used Chasen (Asahara and Matsumoto, 2000) as the Japanese morphological analyzer.

7.2 Data Sets

1. Training Data

   A total of 100,000 Japanese-English simple sentences were used for training data.

2. Development Data

   We used 3,000 sentences for development data for Japanese-English SMT. Of these 3,000 sentences, 375 matched the Japanese-English patterns. We therefore obtained 375 *"English"* sentences for the results. These 375 *"English"* sentences were used as development data for the *"English"*-English translation.

3. Test Data

We used 10,000 Japanese-English simple sentences as test sentences.

7.3 *"English"*-English Phrase Tables

For the second stage, we constructed an *"English"*-English phrase table using Giza++ (Och and Ney, 2003) and "train-model.perl"(Koehn et al., 2007). We set default values for the parameters. Also, 60,000 of the 100,000 training sentences matched the Japanese-English patterns. Thus, we used these 60,000*"English"* sentences to make an *"English"*-English phrase table.

7.4 N-gram Language Model

We built an N-gram language model using 100,000 sentences.

7.5 Decoder

We used "Moses"(Koehn et al., 2007) as a decoder. We also used parameter tuning (MERT) and reordering models. Note that in Japanese-English translation, the position of the verb is sometimes significantly changed from its original position. Thus, we used the unlimited word reordering for a standard SMT. So, we set the "distortion-limit" set to "-1" for a standard SMT. However, our system consists of two-stage machine translation, and the output of the first stage is *"English"*. Consequently, word positions did not dramatically change. Therefore, we set the "distortion-limit" to "6" for the second-stage SMT for our system.

8 Results of our Machine Translation

8.1 Examples of output sentences

Table 10 lists example sentences from our proposed system for the Japanese-English simple sentences. These example sentences are matched with the Japanese-English patterns. In this table, "Input" indicates the input Japanese sentence, "Proposed" indicates the output of our proposed system (PBMT+SMT), "Reference" indicates a correct sentence, and "Moses" indicates the output of a standard SMT.

Table 10: Example Outputs for Japanese-English simple sentences

Input	土手 が 切れ た 。
Proposed	We are out of dikes .
Reference	The bank gave way .
Moses	The bank broke .
Input	この 薬 は 歯痛 に 効く 。
Proposed	This medicine for A toothache .
Reference	This medicine helps a toothache .
Moses	This medicine acts on the toothache .
Input	火 は 台所 から 出 た 。
Proposed	The fire started in the kitchen .
Reference	The fire started in the kitchen .
Moses	The fire started in the kitchen .
Input	内閣 が つぶれる 。
Proposed	The Cabinet collapses .
Reference	The cabinet is dissolved .
Moses	The Cabinet goes bankrupt .
Input	彼女 は フランス へ 行っ た 。
Proposed	She went to France .
Reference	She went over to France .
Moses	She went to France .

8.2 Automatic Evaluations

We used 10,000 test sentences in this experiment. Among these 10,000 sentences, 1,143 sentences matched the Japanese-English patterns. The results of *"English"*-English translation revealed that 725 out of the 1,143 sentences were different compared to the standard SMT system (Moses). The other 8,857 sentences (10,000 - 1,143) did not match the Japanese-English pattern.

We used the BLEU (Papineni et al., 2002) and NIST (NIST, 2003) and METEOR (Banerjee and Lavie, 2005) for evaluation tools. Table 11 summarizes the automatic evaluation results of our machine translation evaluation for the Japanese-English simple sentences. This table shows the results of 1,143 sentences that were matched with the Japanese-English patterns. "Proposed" indicates our proposed system (PBMT+SMT), and "Moses" indicates a standard SMT system.

We obtained a BLEU score of 0.1821 in the Japanese-English simple sentences using our proposed system. In contrast, we obtained a BLEU score of 0.2218 in the Japanese-English simple sentences using the standard SMT system (Moses). This means that our proposed system was not effective for automatic evaluation in the Japanese-English

simple sentences.

Table 11: Experimental Results (1,143 sentence)

	BLEU	NIST	METEOR
Proposed	0.1821	4.817	0.4426
Moses	0.2218	5.239	0.4363

Table 12 shows the all test sentences(10,000 sentences). The 1,143 sentences were translated with the proposed method. The rest of 8,857 sentences were translated with the standard SMT system (Moses).

Table 12: Experimental Results (10,000 sentence)

	BLEU	NIST	METEOR
Proposed	0.1101	4.4511	0.3175
Moses	0.1130	4.5131	0.3160

8.3 Human evaluation

We conducted an ABX test(Clark, 1982), which is a human evaluation method, in order to compare the outputs by the proposed method with those by Moses.

8.3.1 Evaluation Criteria

We organized the outputs into four categories according to the following evaluation criteria. Also, we converted unknown words into the "romaji" characters.

1. Proposed $>$ Moses

   This refers to the case when the output of the proposed method was better than that of Moses. Example sentences are listed in Table 13.

Table 13: Example of "Proposed $>$ Moses"

Input	私 は 彼女 に 結婚 を 申し込ん だ 。
Proposed	I made a proposal of marriage to her .
Reference	I proposed to her .
Moses	I He asked her for her hand .
Input	彼女 は ５ 人 の 子供 を 育て た。
Proposed	She brought up five children .
Reference	She has brought up five children .
Moses	She is five children .

2. Proposed $<$ Moses

   This is the case when the output of Moses was better than that of the proposed method. Example sentences are listed in Table 14.

Table 14: Example of "Proposed $<$ Moses"

Input	仕事 は 山場 に 入っ た 。
Proposed	work went into the labor-management .
Reference	Work has reached the critical point .
Moses	The work is appear to have entered the final stage .
Input	農園 は 道路 に 接し て いる 。
Proposed	The farm is roads are .
Reference	The farm abuts on the road .
Moses	Farm adjoins the road .

3. Proposed $\approx$ Moses

   In this case, the output of the proposed method is the same quality as those by Moses. Example sentences are listed in Table 15.

Table 15: Example of "Proposed $\approx$ Moses"

Input	豊作 に なり そう だ 。
Proposed	It looks like rejoicing .
Reference	The harvest looks promising .
Moses	Hopes looks like .
Input	彼 に よろしく お 伝え ください 。
Proposed	Please send him my best wishes .
Reference	Give him my good wishes .
Moses	Please give my best regards to him .

4. Proposed $=$ Moses

   This refers to when the output of the proposed method and the output of Moses were exactly the same. Examples of such sentences are listed in Table 16.

Table 16: Example of "Proposed $=$ Moses"

Input	彼 は 故郷 を 恋し がっ て いる 。
Proposed	He is homesick .
Reference	He is sick for home .
Moses	He is homesick .

8.3.2 Results of Human Evaluation

We randomly selected 100 sentences from the 1,143 output sentences that were matched with the

Japanese-English patterns. Then we evaluated these 100 sentences. The results are listed in Table 17.

Table 17: Results of Human Evaluation

Proposed $>$ Moses	30 / 100
Proposed $<$ Moses	9 / 100
Proposed $\approx$ Moses	50 / 100
Proposed $=$ Moses	11 / 100

As the table indicates, the proposed method achieved better evaluation than Moses. The p-value was exceeded for 0.95. This means that the proposed method is effective for human evaluation.

9 Discussion

9.1 Analysis of Our Proposed System

Our aim with this system is to reduce the number of ungrammatical sentences produced in machine translation systems. Thus, we analyzed the outputs based on this factor. We compared the output of Moses and the output of our proposed system. And we found that the output of our proposed system affected the output of PBMT, that our system produces more grammatically correct sentences compared to a standard SMT.

9.2 Comparison with Hierarchical phrase-based MT

The pattern acquisition process in the proposed method was similar to the rule extraction of hierarchical statistical phrase-based MT (HSMT). Only, the confident rules are extracted in the proposed method. The reason are discussed follows.

Hierarchical SMT (HSMT) is similar to statistic CFG decoder. So, the number of HSMT parameters is very large. However the number of training data was limited. As the results, they are unreliable and does not perform well, especially for the small amount of training data. Contrast, the proposed method is pattern based. Pattern based approach is similar to network grammar. And it has little parameters compared CFG. So we might obtained these parameters with high reliability.

Also, HSMT has the problem of limiting reordering. The number of spans that are filled during chart decoding is quadratic with respect to sentence length. Hence, it gets worse according as the sentence length increases.

The number of spans that are combined into a span grows linear with sentence length for binary rules, quadratic for trinary rules, and so on. In short, long sentences become a problem. To solved this problem, the size of internal spans has a maximum number. Reordering is limited in hierarchical phrase-based models and should limit reordering for the same reason. On the other hand, the proposed method does not face with such problems because it used patterns. In this reason, we studied the proposed method.

9.3 Improved Pattern Based Statistical Machine Translation

There are many things to improve in PBMT. For example, there is a trade-off between the coverage of input sentences and the translation quality in PBMT. When we made the "high probability phrase table", we set the threshold to 0.1. This was a completely heuristic value. If this value sets low, we obtained many word pairs and many patters. However the reliability of these value was decrease. So we must cut and try this value.

Moreover, there were many bugs in our system. There were 10,000 test sentences in this experiment. Of these 10,000 sentences, 1,143 sentences matched the Japanese-English patterns. We think this number is small for our experience. One possible cause is that we might not have obtained all the possible Japanese-English patterns. We will work on improving the performance of our pattern-based MT system.

10 Conclusion

We developed a two-stage MT system. The first stage consists of an automatically created pattern-based machine translation system. The second stage consists of an phrase-based SMT system. Our goal with this system is to obtain fewer ungrammatical sentences. We performed ABX tests between the output of a standard SMT system (Moses) and the output of the proposed system for 100 sentences. The results indicated that 30 sentences output by the proposed system were evaluated as better than those output by the standard SMT system. In contrast, 9

sentences output by the standard SMT system were thought to be better than those output by the proposed system. This means that our proposed system functioned effectively in the Japanese-English simple sentence task.

We need to overcome several difficulties in order to improve the proposed methods. Moreover, there were many bugs in our system. We will focus on how to solve such difficulties in the future.

References

Masayuki Asahara and Yuji Matsumoto. 2000. Extended models and tools for high-performance part-of-speech tagger. *Proceedings of COLING 2000*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics*.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, M. Laignelet, and F. Rioult. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

David Clark. 1982. High-resolution subjective testing using a double-blind comparator. *J. Audio Eng. Soc*, 30(5):330–338.

Loic Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical postediting on systran's rule-based translation system. *in Second Workshop on SMT*, pages 179–182.

Terumasa Ehara. 2007. Rule based machine translation combined with statistical post editor for japanese to english patent translation. *Proceedings of Machine Translation Summit XI, Workshop on Patent Translation*, pages 13–18.

Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. *Proceedings of the Workshop on Statistical Machine Translation*.

Jin'ichi Murakami, Masato Tokuhisa, and Satoru Ikehara. 2007. Statistical machine translation using large j/e parallel corpus and long phrase tables. *International Workshop on Spoken Language Translation 2007*, pages 151–155.

NIST, editor. 2003. *Automatic Evalation of Machine Translation Quality Using n-gram Co-Occurence Statistics*, http://www.itl.nist.gov/iad/mig/test/mt/.

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

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

Michel Simar, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. *Second Workshop on SMT*, pages 203–206.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. *Intl. Conf. Spoken Language Processing, Denver, Colorado*.

Yushi Xu and Stephanie Seneff. 2008. Two-stage translation: A combined linguistic and statistical machine translation framework. *Proceedings of the Eighth Conference of the Association for Machine Translation*.

Improving Word Alignment by Exploiting Adapted Word Similarity

Septina Dian Larasati
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
SIA TILDE
Riga, Latvia
larasati@ufal.mff.cuni.cz, septina@tilde.lv

Abstract

This paper presents a method to improve a word alignment model in a phrase-based Statistical Machine Translation system for a low-resourced language using a string similarity approach. Our method captures similar words that can be seen as semi-monolingual across languages, such as numbers, named entities, and adapted/loan words. We use several string similarity metrics to measure the monolinguality of the words, such as Longest Common Subsequence Ratio (LCSR), Minimum Edit Distance Ratio (MEDR), and we also use a modified BLEU Score (modBLEU).

Our approach is to add intersecting alignment points for word pairs that are orthographically similar, before applying a word alignment heuristic, to generate a better word alignment.

We demonstrate this approach on Indonesian-to-English translation task, where the languages share many similar words that are poorly aligned given a limited training data. This approach gives a statistically significant improvement by up to 0.66 in terms of BLEU score.

1 Introduction

Low-resourced languages do not have the luxury of having sufficient data to make a good statistical model. In some cases, those languages also do not have any additional language tools to make a linguistically motivated model. This limits the possibilities for low-resourced languages to gain a better translation quality in a Statistical Machine Translation (SMT) experiment.

Word alignment as the basic foundation in phrase-based SMT has gained significant attention in the research community. One of the most commonly applied word alignment approaches in a phrase-based SMT is to combine sets of alignment points learned from two directions (source-to-target and target-to-source). Another approach is to combine different sets of alignment points generated based on different motivations, such as linguistics and heuristics (Xiang et al., 2010). There are also work on using linguistics clues such as string similarity to harvest better word alignments (Bergsma and Kondrak, 2007) or by combining word-level and character-level models in SMT (Nakov and Tiedemann, 2012).

In this paper, we define an algorithm that adds intersecting alignment points on sets of alignment points learned from two different directions. Those added points are points between two similar words (measured by a string similarity metric). Then we apply one of the commonly used word alignment heuristics, MOSES's `grow-diag-final` (*gdfa*), on the new sets of alignment points to generate a better word alignment.

2 The Language Pair

In this work, we choose Indonesian as the low-resourced language and pair it with English. Indonesian-English SMT research is not so prolific. Similar work was done by (Nakov and Ng, 2009) for translating a resource-poor language, Indonesian, to English by using Malay as a pivot language. But most of the related SMT research is done for Malay,

a mutually intelligible language to Indonesian.

Because of the Indonesian complex morphology and the limited data availability, pairing Indonesian and English in an SMT experiment raises a challenge on creating a good word alignment model. Here we try to exploit their orthographically similar word pairs to improve the word alignment.

Some languages that are highly influenced by other languages tend to have similar words. Some of the words may be slightly different in their modified forms. In some cases, we intuitively know how to align words across languages by simply observing their word form.

Although Indonesian has a complex morphology, such as affixation and even reduplication, several Indonesian new words are highly influenced by English, and Indonesian tends to have some loan or adapted words. The words' orthographic similarity can be easily measured, since both languages have the same alphabet. Here we list some word pair examples that we consider semi-monolingual since they are orthographically similar.

Named Entity - Some named entities are poorly aligned because they are scarce in a given limited data. Those named entities both in Indonesian and English have a similar form and can be detected easily, even in their affixed forms, e.g. Indonesian '*Blackberryku*' and the corresponding English 'my Blackberry'.

Loan and Adapted Words - Indonesian adapts several English words and morphemes, e.g.

en	$\leftrightarrow$	***id***
distribution	$\leftrightarrow$	*distribusi*
idealist	$\leftrightarrow$	*idealis*
industry	$\leftrightarrow$	*industri*
department	$\leftrightarrow$	*departemen*
computer	$\leftrightarrow$	*komputer*
president	$\leftrightarrow$	*presiden*

Number and Radix Point - Numbers can come in different combination and are often scarce. They are easy to detect although Indonesian and English radix point are different, where Indonesian uses the comma symbol to separate the integer from the fraction while English uses the dot symbol, e.g. a thousand is 1.000,0 in Indonesian and 1,000.0 in English.

3 Improving the Word Alignment

We improve the word alignment by adding alignment points in the source-to-target (*f2e*) and/or the target-to-source (*e2f*) alignment to add more intersecting alignment points among them. Those intersecting alignment points are added on the word pairs that we consider similar. Then we apply a word alignment heuristic on the new *f2e* and *e2f* sets that now have more intersecting alignment points, to make a new word alignment.

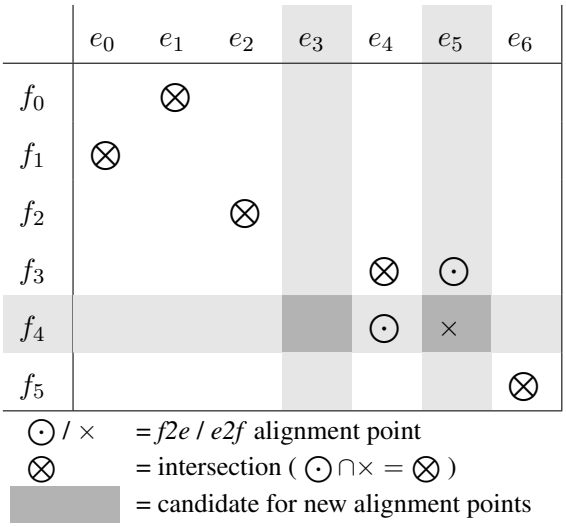

Figure 1: Choosing the candidates for the word pairing and the candidate positions for the new intersecting alignment points.

Suppose $f2e_{ij}$ is a source-to-target alignment link between the *i*-th source word (f_i) and the *j*-th target word (e_j) and $e2f_{ij}$ is a target-to-source alignment between f_i and e_j. Our approach to improve the word alignment is as follow:

1. We choose the source candidate words (*c(f)*) and the target candidate words (*c(e)*), where they are words that are not included in any intersecting alignment points, as illustrated in Figure 1.

2. We pair each *c(f)* to each *c(e)* and score their *string similarity* (*ss*).

3. We choose which pair to be aligned using our *filtering method*.

4. We add alignment points in the *f2e* and/or *e2f* alignment so that the alignment points for the chosen word pair intersect.

5. We apply the `grow-diag-final` (*gdfa*) heuristic[1] on the new *f2e* and *e2f* alignment to produce the new word alignment.

3.1 String Similarity Score

In this work, we uses three different string similarity measures, namely Longest Common Subsequence Ratio (LCSR), Minimum Edit Distance Ratio (MEDR), and a modified BLEU Score (modBLEU). We uses the three metrics to measure our string similarity score (*ss*). Here, we compare the modified BLEU formula to commonly known string similarity metrics, LCSR and MEDR. The LSCR and MEDR formula can be found in Figure 2.

$$ss(f_i, e_j) = LCSR(f_i, e_j) = \frac{|LCS(f_i,e_j)|}{max(|f_i|,|e_j|)} \quad \text{(a)}$$

$$ss(f_i, e_j) = MEDR(f_i, e_j) = 1 - \frac{|MED(f_i,e_j)|}{max(|f_i|,|e_j|)} \quad \text{(b)}$$

Figure 2: The Longest Common Subsequence Ratio (LCSR) and the Minimum Edit Distance Ratio (MEDR) formula for the string similarity metric.

In the modified BLEU, we split the words into characters and we use a modified BLEU on the character level as our string similarity score to measure the characters n-gram precision between the two words.

We score the word pairs using the modified BLEU in two directions: the source word as the hypothesis and the target word as the reference then vice versa. Then we average the two scores. Instead of using at most 4-grams counts in the original BLEU score formula, we modified the formula so that it also consider words with length less than four characters. Below is the formula for the modified BLEU given the length of the hypothesis (*c*) and the length of the reference (*r*).

3.2 Filtering Method

We set an *ss* score threshold to filter the word pairs. We only consider word pairs with a score equal to the threshold and above. We sort the candidate pairs by their *ss* score then by their source token order (*i*) and target token order (*j*). In this way, we pick the most similar word pairs first and then word pairs that occur earlier in the sentence. We assume that the similar words have the same order of occurrence in the sentence.

$$BLEU_m(f_i, e_j) = BP \bullet exp(\textstyle\sum_{n=1}^{chk} log(p_n)) \quad (1)$$

$$BP = min(1, e^{1-r/c}) \quad (2)$$

$$chk = min(4, r, c) \quad (3)$$

$$ss = (BLEU_m(f_i, e_j) + BLEU_m(e_j, f_i))/2 \quad (4)$$

Figure 3: The modified BLEU formula for the string similarity metric.

All the newly added alignment points have to be one-to-one aligned. We discard any new alignment point that violates this condition, as we pick the word pairs.

Consider Figure 1, if f_4, e_3, and e_5 all are the same word '*street*', so that $BLEU_m(f_4, e_3) = BLEU_m(f_4, e_5) = 100\%$, only a link between f_4 and e_3 is added, because the pair occurs earlier in the sentence and adding another link between f_4 and e_5 will violates the one-to-one alignment. If f_4, e_3, and e_5 are '*street*', '*streen*', '*street*' respectively, only a link between f_4 and e_5 is added, because it is chosen first for its better score.

4 Experiment

4.1 Data

The corpus we use in this work is the IDENTIC (Larasati, 2012) Indonesian-English parallel corpus. Our training, tuning, and testing data contain around 43K, 1K, and 1K parallel sentences respectively. The sentences are taken randomly without replacement from the corpus.

4.2 Common Setting

The SMT system is in lowercased-to-lowercased Indonesian-to-English translation direction. We use the state-of-the-art phrase-based SMT system MOSES (Koehn et al., 2007). We use GIZA++ tool

[1] http://www.statmt.org/moses/?n=FactoredTraining.AlignWords

(Och and Ney, 2003) to build the bidirectional sets of alignment points (*f2e* and *e2f*).

For the *baseline* system, we run the MOSES *gdfa* heuristic on the initial *f2e* and *e2f*. And for the experiment systems, we apply our algorithm to the initial *f2e* and *e2f* to generate the new sets and then we apply the same *gdfa* heuristic on the new sets. This makes the *gdfa* heuristic algorithm starts with more intersecting alignment points.

We create the English Language Model (LM) using SRILM (Stolcke, 2002) on the English Europarl corpus. The quality of the translation results are measured using BLEU score (Papineni et al., 2002) and pairwise bootstrapping significance test (Koehn, 2004).

4.3 Result

We set up the *baseline* system and the *exact* system. Then we created five experimental SMT systems that are set with different *ss* score thresholds, namely 90, 80, 70, 60, and 50.

The *exact* system aligns word pairs that are orthographically equal. Here the algorithm successfully aligns the foreign words ('supreme', 'court', etc), named entities ('wall' 'street', 'telkom', 'jakarta'), numbers ('1.28', '4.1', etc), and punctuations.

As we use different thresholds, the algorithm can pair Indonesian affixed words ('*uraniumnya*' with 'uranium'), adapted words ('*internasional*' with 'international' or '*kwartet*' and 'quartet'), and different number formatting ('0,85863' with '0.85863'). It also captures and pairs some inconsistent number formatting such as '6.5' and '6.50' and some misspelling such as '*streen*' and 'street' of the word 'wall street'.

In general, the translation quality increases when we add links for very similar words measured by any of the string similarity metrics. When we use the modified BLEU metric, the system's BLEU score is increasing in a logarithmic scale when the threshold is between 60 and 100.

But as the threshold set to a lower value, it wrongly aligns some short stopwords such as the Indonesian '*ini*' (this) with the English preposition 'in' and the Indonesian '*itu*' (that) with the English personal pronoun 'it', which makes the translation quality become poor. When we use the LCSR and MEDR metric, the translation quality decreases earlier with a bigger threshold.

	System	Δ to *exact*	BLEU
	baseline	-9469	27.25
	exact	0	*27.62
modBLEU	th*ss*-90	16	**27.83
	th*ss*-80	390	**27.87
	th*ss*-70	857	**27.89
	th*ss*-60	1457	**27.91
	th*ss*-50	3543	27.23
LCSR	th*ss*-90	107	**27.82
	th*ss*-80	1321	*27.52
	th*ss*-70	2813	27.42
	th*ss*-60	8112	27.12
	th*ss*-50	30214	*27.40
MEDR	th*ss*-90	83	*27.74
	th*ss*-80	975	*27.70
	th*ss*-70	2185	**28.03
	th*ss*-60	5693	27.25
	th*ss*-50	18037	27.12

* / **) 90% / 95% statistically significant

Table 1: SMT systems evaluation in term of BLEU score. The experiment systems with different thresholds are named th*ss*-[*threshold*]. Δ to *exact* is the number of the added intersection points compared to the *exact* system.

Table 1 summarizes the number of the added intersecting alignment points and the evaluation for the *baseline* and the experiment systems.

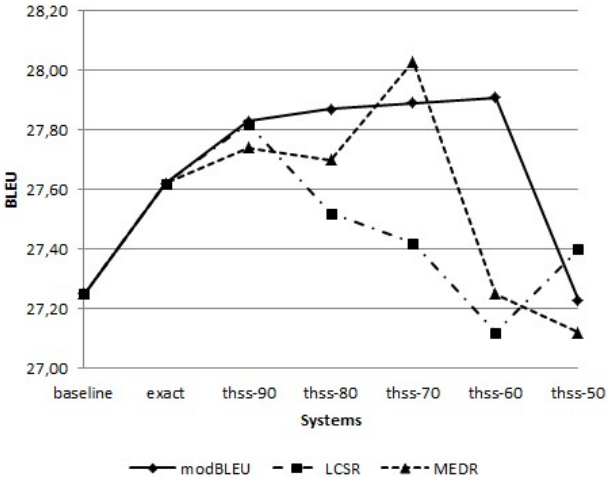

Figure 4: The *baseline* and the experimental SMT systems translation quality in terms of BLEU Score.

5 Conclusion

Our method captured similar words that are semi-monolingual across languages, such as numbers, named entities, and adapted words. We used this information as a clue to add alignment points. We showed that adding good quality intersecting alignment points before applying the *gdfa* heuristic helps to gain a better translation quality for a Indonesian-to-English SMT system. We used LCSR and MEDR as string similarity metrics and we also introduced another metric, a modified BLEU formula. We still found some word pairs that are wrongly aligned and most of them are stopwords. Modifying the string similarity formula or the filtering method so that it does not capture these stopwords will be a good future improvement.

Acknowledgments

The research leading to these results has received funding from the European Commission's 7th Framework Program under grant agreement n° 238405 (CLARA), by the grant LC536 Centrum Komputační Lingvistiky of the Czech Ministry of Education, and this work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

References

Shane Bergsma and Grzegorz Kondrak. 2007. Alignment-based discriminative string similarity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 656–663, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Septina Dian Larasati. 2012. IDENTIC corpus: Morphologically enriched indonesian-english parallel corpus. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1367, Singapore, August. Association for Computational Linguistics.

Preslav Nakov and Hwee Tou Ng. 2011. Translating from morphologically complex languages: a paraphrase-based approach. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL2011), Portland, Oregon, USA*.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–305, Jeju Island, Korea, July. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*.

Bing Xiang, Yonggang Deng, and Bowen Zhou. 2010. Diversify and combine: Improving word alignment for machine translation on low-resource languages. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 22–26, Uppsala, Sweden, July. Association for Computational Linguistics.

Addressing some Issues of Data Sparsity towards Improving English-Manipuri SMT using Morphological Information

Thoudam Doren Singh
Centre for Development of Advanced Computing (CDAC)
Gulmohor, Cross Road 9
Juhu, Mumbai-400049, India
thoudam.doren@gmail.com

Abstract

The performance of an SMT system heavily depends on the availability of large parallel corpora. Unavailability of these resources in the required amount for many language pair is a challenging issue. The required size of the resource involving morphologically rich and highly agglutinative language is essentially much more for the SMT systems. This paper investigates on some of the issues on enriching the resource for this kind of languages. Handling of inflectional and derivational morphemes of the morphologically rich target language plays important role in the enrichment process. Mapping from the source to the target side is carried out for the English-Manipuri SMT task using factored model. The SMT system developed shows improvement in the performance both in terms of the automatic scoring and subjective evaluation over the baseline system.

1 Introduction

Since the dawn of SMT system in the early 90s with the seminal work at IBM (Brown et al., 1992; Brown et al., 1993), there has been growth in the number of the SMT system for several language pairs. Performant SMT systems for the major languages are available. In the same time, such development is limited for less privileged and resource poor languages. Developing English-Manipuri SMT systems is one of such examples. Manipuri is a morphologically rich and highly agglutinative in nature. New words are easily coined by combination of various morphemes. Verb morphology is more complex and productive than noun morphology. In Manipuri, adjective and adverbs come from verbal root through derivational morphology. Aspectual marker goes with the derived forms. This language contains abundant reduplicated multiword expressions (RMWE). Language resource for this language pair is not available in the required measure.

2 Related Work

Several SMT systems between English and morphologically rich languages are reported. (Oflazer and El-Kahlout, 2007) investigated different representational granularities for sub-lexical representation in statistical machine translation work from English to Turkish by exploring different representational units in English to Turkish SMT. (Yeniterzi and Oflazer, 2010) further reported syntax-to-morphology mapping in factored phrase-based Statistical Machine Translation (Koehn and Hoang, 2007) from English to Turkish relying on syntactic analysis on the source side (English) and then encodes a wide variety of local and non-local syntactic structures as complex structural tags which appear as additional factors in the training data. On the target side (Turkish), they only perform morphological analysis and disambiguation but treat the complete complex morphological tag as a factor, instead of separating morphemes. Some of the SMT systems between English and morphologically rich languages which used morphemes to address the data sparsity are discussed below.

English-to-Czech phrase-based machine translation experiment (Bojar, 2007) with additional annotation of input and output tokens

(multiple factors) used to explicitly model morphology by setting up various multiple factors and the amount of information in the morphological tags resulted in significant translation quality increase. Further, two contributions using factored phrase based model and a probabilistic tree transfer mode at deep syntactic layer are made by (Bojar and Hajič, 2008) of English-to-Czech SMT system. (Toutonova et al., 2007) reported the improvement of an SMT by applying word form prediction models from a stem using extensive morphological and syntactic information from source and target languages. (Gandhe et al., 2011) proposed a solution to augment the phrase table with all possible forms of a verb for improving the overall accuracy of the English–Hindi MT system by using simple stemmer and easily available monolingual data to generate new phrase table entries that cover the different variations seen for a verb. (Habash, 2008) presented four techniques for online handling of Out-of-Vocabulary words in Arabic-English Phrase based Statistical Machine Translation by using spelling expansion, morphological expansion, dictionary term expansion and proper name transliteration to reuse or extend a phrase table.

3 Enriching the Language Resource

SMT systems demand a large parallel corpus as training data. One of the approaches to develop SMT systems for less privileged and resource poor language is to enrich the resource through morphological processing by learning the general rules of morphology. This helps to increase the training data size and increase the coverage of the occurrence of the different words. The representation of words, i.e. the spelling has the most important role to play. Including different forms of a word in the training data makes a sense to improve the translation quality. While exploring the different representational units from English-to-Manipuri, there is lack of information at the source side for derivation and inflection of the target words. The verb morphology is more complex and productive. Focusing on the verb, the derivational morphology is more productive than the inflectional. Separating lemma, inflections and derivational morphemes allows the system to learn more about the different possible word formations. Thus, considering all the possible morpheme combination helps to enrich the language in order to cover more vocabulary.

3.1 Examples of Word Level Alignment between English and Manipuri

Consider the following example of English to Manipuri translation depicting the word level alignment by figure 1, stems and separated affixes by Figure 2 and phrasal level alignment by figure 3. From the example we can see that for each Manipuri word, there is a corresponding chunk of English (i.e., a group of words). This prompts us to correlate a group of English words with the help of a chunker and a Manipuri word.

But the question is to sort out how many possible morphemes are there to represent a word. Consider the following Manipuri sentence showing the possible translations due to non-standardized spelling.

English:

They came.

Manipuri Translation: (different possible orthographic and phonographic variations)

মখোয় লাক্লম্মী ।	(*makhoy lak-lammee*)
মখোয় লাকঅম্মী ।	(*makhoy lak-ammee*)
মখোয় লাক্লমঈ ।	(*makhoy lak-lam-ee*)
মখোয় লাক্লমই ।	(*makhoy lak-lam-i*)
মখোই লাক্লমই ।	(*makhoi lak-lam-i*)

Enrichment of the language resource is carried out for this languages category using morphemes with various orthographic variations. Inclusion of these various orthographic variations is absolutely necessary. However, the problem crops up how much of the training data is really increased in terms of size and improved in terms of alignment quality? When they are not handled in one of these steps these words become out-of-vocabulary (OOV) words. So, spelling expansion of these words in order to extend the phrase table becomes essential. This helps in the orthographic normalization and minimizing the data sparsity. The spelling expansion and morphological expansion are definitely helping minimizing data sparsity.

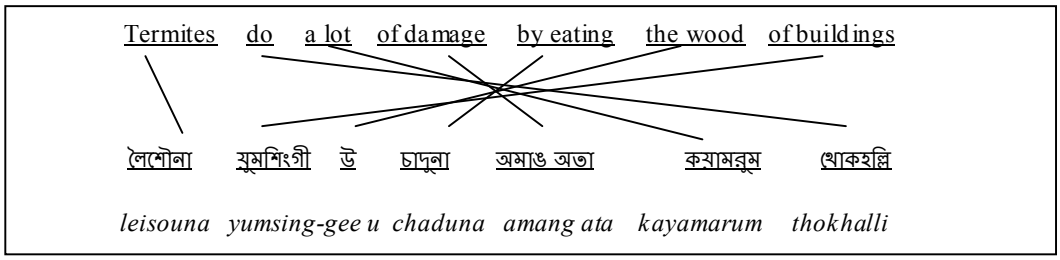

Figure 1: Word and Chunk level alignment between English and Manipuri

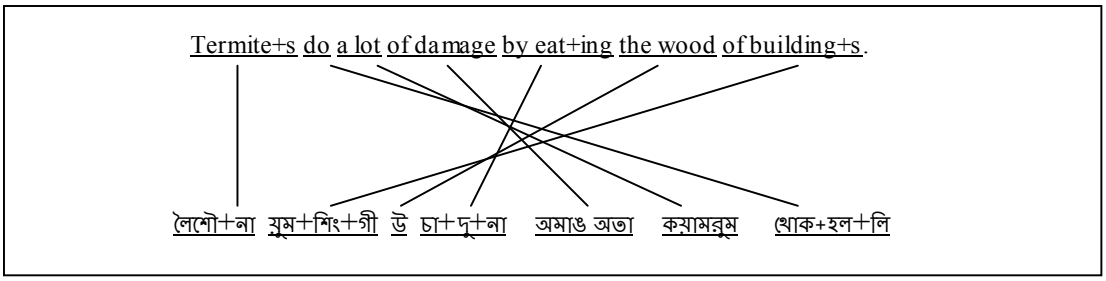

Figure 2: Stems and Morphemes separated between English and Manipuri alignment

(NP (NNS Termites)) → লৈশৌ+না
(VP (VBP do) → থোকহল+লি
(NP (DT a) (NN lot)) → কয়ামরুম
(PP (IN of)
(NP (NN damage))) → অমাঙ অতা
(PP (IN by)
(VP (VBG eating) → চা+দু+না
(NP (DT the) (NN wood)) → উ
(PP (IN of)
(NP (NNS buildings))) → য়ুম+শিং+গী

Figure 3: Phrasal Level Alignment between English and Manipuri

4 Key Aspects of Manipuri Morphology

In this agglutinative language the numbers of verbal suffixes are more than that of the nominal suffixes (Singh, 2000). New words are easily formed in Manipuri using morphological rules. There are 8 inflectional (INFL) suffixes and 23 enclitics (ENC). There are 5 derivational prefixes out of which 2 are category changing and 3 are non-category changing. There are 31 non-category changing derivational suffixes and 2 category changing suffixes. The non-category changing derivational suffixes may be divided into first level derivatives (1st LD) of 8 suffixes, second level derivatives (2nd LD) of 16 suffixes and third level derivatives (3rd LD) of 7 suffixes. Enclitics in Manipuri fall in six categories: determiners, case markers, the copula, mood markers, inclusive/exclusive and pragmatic peak markers and attitude markers. The categories are determined on the basis of position in the word (category 1 occurs before category 2, category 2 occurs before category 3 and so on). Manipuri morphological processing works are reported by (Singh and Bandyopadhyay, 2006) and (Singh and Bandyopadhyay, 2008).

4.1 Verb morphology

Three derivational categories may optionally precede the final inflectional suffix. The 1st LD suffixes signal adverbial meanings, the 2nd LD suffixes indicate evidentiality, the deitic reference of a verb, or the number of persons performing the action and the 3rd LD suffixes signal aspect and mood. Verb roots may also be used to form verbal nouns, adjectives and adverbs. Verbal nouns are formed through the suffixation of the nominalizer পা –pə to the verb root. The following is the list of word structure rules for verbs (Shobhana, 1997)

a. Verb → STEM INFL
b. STEM → Stem (3rd LD)
c. Stem → Stem (2nd LD)
d. Stem → Root (1st LD)
e. ROOT → root (root)
f. 3rd LD → (mood1)(mood2)(aspect)
g. 2nd LD → (2nd LD1),(2nd LD2),(2nd LD3)..
h. 1st LD → 1st LD

Derivational Prefixation	Root	1st LD	2nd LD	3rd LD	Inflection

Figure 4: General form of Verb Morphology

There are 3 categories (mood1, mood2, and aspect) belonging to the third level derivational (3rd LD) markers. The general form of verb morphology is shown in figure 4.

The sub-categorization frames of affixes will restrict that only nominal affixes occur with a noun and verbal affixes occur with a verb root. The derivational suffix order of the word চেকখাইরক্কনি (It'll get cracked) is given below:-

চেক	খাই.	রক	ক	নি
cek	*–khay*	*-rək*	*-kə*	*-ni*
crack	*-totally affect (1st LD)*	*-distal (2nd LD)*	*-potential (3rd LD)*	*–copula*

The রক *-rək* has allomorph লক-*lək*. রক *-rək* occurs after vowels while লক-*lək* occurs after consonants,

চারকএ *–ca-rək-ey* (ate there and came here)
চাল্লকএ *– cam-lək-ey* (washed there and came here)

The formation of verb can be of the form

Verb stem + aspect/mood → verb
থক *-thək* (drink) + লে *-le-* → থক্কলে *-thəkle* (has drunk)

The verbal noun is formed with the rule as given as

Verb Stem + Nominalizer → Verbal noun
থোং *-thong* (cook)+ বা *-ba* → থোংবা *-thongba* (to cook)

4.2 Noun Morphology

The following is the list of word structure rules for nouns (Shobhana, 1997)

N → Stem INFL (ENC)
Stem → stem (2nd LD)
Stem → ROOT(1st LD)
ROOT → (prefix) root (root)

Figure 5 shows the general form of noun morphology in Manipuri. Examples of some singular/plural noun forms are listed in Figure 6.

Pronominal prefix	Root	gender	number	Quantifier	Case

Figure 5: General form of Noun Morphology

Singular Form	Plural Form
উচেক -Uchek (bird)	উচেকশিং -Ucheksing(birds)
ম -Ma (He/She)	মখোয় -Makhoy (they)
মী -Mi (man)	মীয়াম -Mi-yaam (men)

Figure 6: Singular/Plural forms

Although case markers are functionally inflectional, they exhibit the clitic like characteristic of docking at the edge of a phrase. The word structure of rules of verbs and nouns are identical except for the category of the word level node, the possible terminal elements of the derivational and inflectional categories and the lack of the third level nominal derivation. Two examples to demonstrate the noun morphology are given below:-

মচানুপীশিংনা *(mə-ca-nu-pi-siŋ-nə)* 'by his/her daughters'
মচানুপাশিংনা (*mə-ca-nu-pa-siŋ- nə*) 'by his/her sons'

The ম *-mə* 'his/her' is the pronominal suffix and চা *-ca* 'child' is the noun root. The নু *-nu* 'human' is suffixed by পী *-pi* to indicate a female human and পা–*pa* to indicate a male human. শিং –*siŋ* or খোই *-khoy* or য়াম–yaam can be used to indicate plurality. *-siŋ* cannot be used with pronouns or proper nouns and *-khoy* cannot be used with nonhuman nouns. না *-nə* meaning 'by the' is the instrumental case marker.

4.3 Adjectives and Adverbs

In Manipuri, adjective and adverbs come from verbal root (in the example: ফ (ph)) through derivational morphology. Aspectual marker goes with the derived forms. Some of the examples are given as:

a)
The player is good
শান্নরোয়দু ফৈ
shannaroydu phei

b)
The player is still good
শান্নরোয়দু হৌজিকসু ফরি
shannaroydu houjiksu phari

c)
The player is always good
শান্নরোয়দু অদুম ফৈ
shannaroydu adum phei

d)
The player was good
শান্নরোয়দু ফরম্মী
shannaroydu pharami

5 Experiments

Indian languages are morphologically rich and have relatively free-word order where the grammatical role of content words is largely determined by their case markers and not just by their positions in the sentence. SMT systems between English and morphologically rich and highly agglutinative languages suffer badly if adequate training and language resource is not available and the accountability of individual morpheme is not considered. Machine Translation systems of Manipuri (the first Tibeto-Burman language for which MT system is developed) and English are reported by (Singh and Bandyopadhyay, 2010b) on developing the first Manipuri to English example based machine translation system followed by (Singh and Bandyopadhyay, 2010c) on development of English-Manipuri SMT system using morpho-syntactic and semantic information where the target case markers are generated based on the suffixes and semantic relations of the source sentence. Further (Singh and Bandyopadhyay, 2011a) reported on the development of bidirectional SMT system for English-Manipuri language pair using dependency relations, morphological information and parts of speech tags and (Singh and Bandyopadhyay, 2011b) continued reporting on the integration of reduplicated multiword expression and named entities into the English-Manipuri SMT system. In the present work, detailed morphological information such as the inflection and derivational morphemes are integrated into the system. In an effort to optimize, the training data, we experimented on the variation in the performance while making choice of sentence length for training. Being a highly agglutinative language, the translation performance is largely affected for longer training. The English-Manipuri parallel corpus on news domain developed by (Singh and Bandyopadhyay, 2010a) is used in the experiment. Two different models are developed as shown in figure 7; (i) inflectional model and (ii) inflectional + derivational model. The inflectional model uses lemma and suffix factors on the source side, lemma and suffix on the target side for lemma to lemma and suffix to suffix translations with generation step of lemma plus suffix to surface form. For the second model, two important treatments of the noun (Thoudam, 1982) and verb morphology for mapping with the corresponding source side is carried out considering the stem, derivational morphology and inflectional morphology of the target side. The reason why Manipuri inflectional morphology is to be treated as separate factor is

that – it is comparatively easier to map to English dependency relations and suffix information (Singh and Bandyopadhyay, 2010c) to address the crux of the fluency. The BLEU score of this SMT system on the same corpus statistics as given in table 1 is 16.873 as reported earlier. Again, considering all Manipuri derivational morphology as one factor mapped to the complete phrase of the corresponding English phrases helps to cover overall meaning. This exercise reduces the overall burden to deal with individual morphemes with its discontinuous representation of English counterpart. We augment the training phrase table with 5000 manually prepared variants of verbs and nouns phrases for improving the overall accuracy of the SMT system. Manipuri uses Bengali script to represent the text. The wide variations of tone are not captured during the textual representation. Lexical ambiguity is very common in this language. This has resulted towards the requirement of a word sense disambiguation module. As part of this ongoing experiment, an additional lexicon of 11000 entries between English and Manipuri is employed to handle bits of sense disambiguation with the help of a word-based language model. The English sentences are processed with morpha (Minnen, 2003) and Stanford Parser (de Marneffe and Manning, 2008) is used for parsing. All words in Manipuri are bound except noun. We process the Manipuri corpus by segmenting into three parts, viz, (a) lemma (b) derivational morphemes and (c) inflections. The Manipuri stemmer (Singh and Bandyopadhyay, 2008) is used to separate the stem, derivational morpheme and inflections.

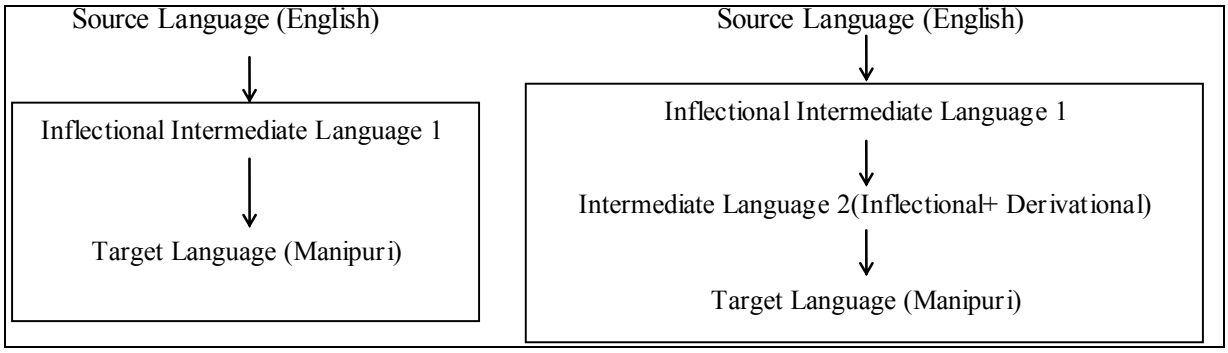

Figure 7: Two different models of English-Manipuri translation

For word-based language modeling, 200,000 Manipuri news sentences are used. The approach is to take account of the various function words by grouping after syntactic analysis of the source side sentences using Stanford Parser (de Marneffe and Manning, 2008) and map to the corresponding target side of Manipuri after morphological analysis. The baseline system is based on the surface word forms using the Moses (Koehn et al., 2007) default setting. A factored translation model can embed multiples of monolingual MT systems inside of it. So, the addition of 'derivational morphology' has similar effect that one additional intermediate language includes the factored form. Consider the sentence 'I should have done it' meaning 'ঐ মদু তৌরমগদবনি।',(*ei madu touramgadabanee*) the phrase 'should have done' meaning তৌরমগদবনি consist of the inflections as well as derivational morphemes. The meaning of 'should have' is the derivational morpheme i.e., রমগদব and নি is the copula. The inflectional model gives the output as 'ঐ মদু তৌনি' meaning 'I'll do it'. But, considering the inflectional + derivational model, the output is adequately addressed by the induction of the syntactic information from the source side. The mapping from the auxiliary verbs to the derivational morphemes can help to improve as an important factor.

The various models developed are evaluated using BLEU (Papineni et al., 2002) and (NIST Doddington, 2002) automatic scoring techniques. SRILM is used for language modeling (Stolcke, 2002).

1	**English:** The number of teachers required is 58 for arts subjects and 32 for science subjects. Reference: অৱাংপা ওজা মশিং আর্টস সবজেক্তনা ৫৮ অমসুং সাইন্স সবজেক্তনা ৩২ নি Baseline: অৱাংপা ওজা মশিং খাকি সাইন্সকি ৩২ আর্টসতা ৫৮ Inflectional: অৱাংপা ওজা মশিং খাকি সাইন্সকি ৩২ আর্টসতা ৫৮ নি Derivational+ Inflectional: সবজেকশিং অৱাংপা ওজা মশিং আর্টসনা ৫৮ অমসুং সাইন্সনা ৩২ নি	Most of the meaning is conveyed by all the models.
2	**English:** The branches of the tree spread out in all directions. Reference: উগী মশা ময়ামদুনা মায়কৈ খুদীংদা লোঙখোকই Baseline: হৌবা উ খন্দোকপা মায়কৈ Inflectional: উগী মশা লোঙখোকই Derivational + Inflectional: উগী মশাশিং মায়কৈ লোঙখোকই	Poor meaning is conveyed by all the models.
3	**English:** Termites do a lot of damage by eating the wood of buildings. Reference: লৈশৌনা য়ুমগী উ চাখোক্লুনা অমাঙঅতা কয়ামরুম খোকহল্লি Baseline: লৈশৌনা তৌবসি য়াল্লা মাঙহনবা মীওইশিংনা ময়ুমদগী উ Inflectional: লৈশৌনা মাঙহনবা মীওইশিংনা ময়ুম উ Derivational + Inflectional : লৈশৌনা য়ুমগী উ চাখোক্লুনা অমাঙবা খোকহল্লি	No meaning is conveyed by all the models except the output of Derivational + Inflectional model.
4	**English:** Contract works should be given only to those who would be able to carry out the work with sincerity and dedication. Reference: খবক নিংথিনা তৌগদবা মীদা ঠিকা পীগদবনি Baseline: কন্ট্রেক্ট ৱার্ক পীগদবনি ঙসিগী খুল্লাই অসিদা carrying হায়বসি শেংনা খবক গ্রুপশিংনা Inflectional: কন্ট্রেক্ট ৱার্ক শেংনা খবক পীগদবনি Derivational + Inflectional: কন্ট্রেক্ট ৱার্ক খবক গ্রুপশিংনা হায়বসি পীগদবনি	Poor meaning is conveyed by all the models.
5	**English:** The election office has reportedly intimidated the IFCD for taking up necessary measures. Reference: ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা খবক পায়খল্লবা খঙহনখ্রে Baseline: ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা খবক পায়খল্লবা খঙহনখ্রে Inflectional: ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা খবক পায়খল্লবা খঙহনখ্রে Derivational + inflectional : ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা খবক পায়খল্লবা খঙহনখ্রে	Full meaning is conveyed by all the models.

Figure 8: Output of various models

	Number of sentences	Number of words
Training	10350	296728
Development	600	16520
Test	500	15204

Table 1: Corpus Statistics

	BLEU	NIST
Baseline	13.045	4.25
Inflectional	15.237	4.79
Derivational+ Inflectional	15.824	4.85

Table 2: Automatic Evaluation Scores

Table 1 show the corpus statistics and table 2 shows the automatic evaluation scores. The incorporation of derivational morphemes improves the BLEU and NIST scores by capturing a larger coverage of word forms.

6 Conclusion and Future Direction

With the present work, we have identified a novel approach to integrate finer linguistic details into the translation model by taking into account of the syntactic information from the source side and derivational and inflectional morphemes from the target side. Though English-Manipuri parallel corpora is limited in size, the performance of the SMT system is improved by taking into account of the above mentioned morphemes and thus helping to address the data sparsity problem for developing SMT systems between English and highly agglutinative and morphologically rich language like Manipuri. Our stress is mainly on the noun and verb morphology. Figure 8 shows the variations in the output of different models based on subjective evaluation. The automatic evaluation metrics shows the improvement of the scores for translation models catering more linguistic morphemes than the baseline models. The scalability of the present task is to develop SMT system between English and morphologically rich languages but with limited parallel corpora.

Acknowledgments

I, sincerely, thank Dr. Zia Saquib, Executive Director, CDAC (Mumbai), Prof. Sivaji Bandyopadhyay, Jadavpur University, Kolkata and the two anonymous reviewers for their support and valuable comments.

References

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In Proceedings of the Intl. Conf. on Spoken Language Processing.

Ankur Gandhe, Rashmi Gangadharaiah, Kartik Vishweswariah and Ananthkrishnan Ramanathan. 2011. Handling Verb Phrase Morphology in Highly Inflected Indian Languages for Machine Translation, In proceedings of the 5th International Joint Conference on Natural Language Processing, Pages 111-119, Chiang Mai, Thailand.

Ch. Yashawanta Singh. 2000. Manipuri Grammar. Rajesh Publications, New Delhi.

George Doddington. 2002. Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics. In Proceedings of HLT 2002, San Diego, CA.

Guido Minnen, John Carroll and Darren Pearce, 2001. Applied Morphological Processing of English, Natural Language Engineering, 7(3), pages 207-223

Kemal Oflazer, and I. Durgar El-Kahlout. 2007. Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation, in Proc. of the 2nd Workshop on Statistical Machine Translation, pages 25–32.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of 40th ACL, Philadelphia, PA.

Kristina Toutanova, Hisami Suzuki and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation, in Proc. 46th Annual Meeting of the Association for Computational Linguistics.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford Typed Dependency Manual.

Nizar Habash. 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation, In

Proceedings of Association for Computational Linguistics-08.

Ondřej Bojar. 2007. English-to-Czech FactoredMachine Translation. In Proc. of ACL Workshop on Statistical Machine Translation, pages 232–239, Prague.

Ondřej Bojar and Jan Hajič. 2008. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation, Proceedings of the Third Workshop on Statistical Machine Translation, pages 143–146, Columbus, Ohio, USA.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John D. Lafferty and Robert L. Mercer, 1992. Analysis, Statistical Transfer, and Synthesis in Machine Translation. In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, pages 83-100, Montreal, Canada.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer, 1993. Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, pages 163-311.

Purna C. Thoudam. 1982. Nouns in Meiteiron, Linguistics of the Tibeto Burman Area, Vol 6.2, Spring 1982. http://sealang.net/sala/archives/pdf8/thoudam1981 nouns.pdf.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models, Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish, In proceeding of the 48th Annual Meeting of the Association of Computational Linguistics, Pages 454-464, Uppsala, Sweden.

Shobhana L. Chelliah. 1997. A Grammar of Meithei. Mouton de Gruyter, Berlin, pages 77-92.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2006. Word Class and Sentence Type Identification in Manipuri Morphological Analyzer, Proceeding of MSPIL 2006, IIT Bombay, pages 11-17, Mumbai, India.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2008. Morphology Driven Manipuri POS Tagger, In proceedings IJCNLP-08 Workshop on NLPLPL, pages 91-98, Hyderabad, India.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010a. Semi Automatic Parallel Corpora Extraction from Comparable News Corpora, In International Journal of POLIBITS, Issue 41 (January – June 2010), ISSN 1870-9044, pages 11-17.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010b. Manipuri-English Example Based Machine Translation System, International Journal of Computational Linguistics and Applications (IJCLA), ISSN 0976-0962, pages 147-158

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010c. Statistical Machine Translation of English-Manipuri using Morpho-Syntactic and Semantic Information, In proceedings of Ninth Conference of the Association for Machine Translation in Americas (AMTA 2010), pages 333-340, Denver, Colorado, USA.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2011a. Bidirectional Statistical Machine Translation of Manipuri English Language Pair using Morpho-Syntactic and Dependency Relations, In International Journal of Translation, Vol. 23, No.1 (Jan-Jun), 2011, pages 115-137.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2011b. Integration of Reduplicated Multiword Expressions and Named Entities in a Phrase Based Statistical Machine Translation System, Proceedings of the 5th IJCNLP, pages 1304–1312, Chiang Mai, Thailand.

Statistical Machine Translation for Depassivizing German Part-of-speech Sequences

Benjamin Gottesman
Acrolinx
Friedrichstraße 100
10117 Berlin, Germany
ben.gottesman@acrolinx.com

Abstract

We aim to use statistical machine translation technology to correct grammar errors and style issues in monolingual text. Here, as a feasibility test, we focus on depassivization in German and we abstract from surface forms to parts of speech. Our results are not yet satisfactory but yield useful insights into directions for improvement.

1 Introduction

There exist software applications that identify errors in text and, in some instances, automatically generate possible corrections. However, they are not yet sophisticated enough to correct most grammar errors and style issues. We aim to train instances of the statistical machine translation (SMT) system Moses (Koehn et al, 2007) to perform monolingual reformulations aimed at correcting specific grammar and style issues. In other words, the system will 'translate' from error-filled text to correct text in the same language. If successful, this approach could be used to extend the capacity of existing text-checking software to generate corrections.

This work is at an early stage. What we present here is a feasibility test that is limited to a specific language and style issue: avoiding passive voice in German. Furthermore, we abstract from surface forms to parts of speech (POSes), thereby admittedly glossing over some complications, as we will show.

2 Related Work

This is related to earlier work on statistical post-editing. Dugast et al (2007), for example, trained Moses on a parallel corpus that paired the outputs of SYSTRAN, a rule-based machine translation (RBMT) system, with gold-standard human translations of the same input sentences. That is, they trained an SMT system to correct RBMT errors. Our training data, in contrast, pairs human-written sentences containing errors with human-edited versions of the same sentences in which the errors have been corrected.

This is also related to work in which POS information is used in SMT. Popović and Ney (2007) and Genzel (2010), for example, perform POS-based reordering on source-language sentences in order to make them more like the target language and thereby reduce the amount of reordering that must be performed by the SMT system. Stated more generally, they use POS information in a preprocessing step that makes the SMT task easier.

We, in contrast, provide the POSes as input to the SMT system itself, which makes our work more similar to work in the area of factored translation models. Koehn and Hoang (2007) presented these models and described experiments in which they were used as a means of annotating SMT training data with POSes and other lexical information. We do not technically use factored models, though, as we train our SMT systems on the POSes alone rather than on surface forms annotated with POSes.

3 Experiment Method and Results

Concisely, our method is as follows.

- check German text segments using Acrolinx
- manually correct issues identified by Acrolinx flags
- select segment pairs (before and after editing) in which the only edit was a depassivization
- convert segments to POS-tag sequences
- where two segment pairs are duplicates at POS-tag level, discard one
- partition segments into training/tuning/test sets
- train and tune SMT depassivizer and apply it to test set
- evaluate test output automatically and manually

The following subsections describe the method in more detail.

3.1 Data

We started from approximately 77,000 German text segments from the OPUS corpus (Tiedemann, 2012). This consists of technical documentation of the OpenOffice office productivity software suite. We checked these segments using Acrolinx, a commercial text-checking software product that flags spelling and grammar errors as well as style issues (as described by Bredenkamp et al (2000)). We then had a human editor edit the segments in response to the flags. The editor was not permitted to perform any edit that was not in response to a specific flag, but was permitted to ignore false flags or other flags for which there was no useful edit. For each edited segment, the editor noted which flag type(s) prompted the edit(s). Thus, we can train an SMT system to correct one specific type of error by selecting as training data just those segments containing relevant edits. Table 1 shows the number of edits for the most common flag types.

The most common of all is *Avoid passives*, a flag that reflects the stylistic dispreference for passive voice in, for example, German technical writing (tekom, 2011). We focus on this flag type because of the amount of data and because its correction patterns are largely systematic yet not implemented by existing text-checking software. Thus, we shall try to train Moses to depassivize German sentences. We use only the isolated edits in order to avoid confusing the system with unrelated edits. The available data is thus 571 German OpenOffice text segments, before and after depassivization by a human editor, with no other edits. We use a POS tagger to convert the text segments to sequences of POS tags. After removing some duplicates among the segments, we partition the remainder arbitrarily into training, tuning, and test sets of 517, 20, and 10 POS-tag sequence pairs, respectively.

Flag type	# edited segments	# isolated* edits
Avoid passives	1411	571
Avoid ambiguous words	611	291
Avoid parentheses	554	393
Avoid more than two prepositional phrases	504	140
Use digits	347	178
Avoid pronouns with unclear referent	340	90
Avoid verbosity	331	121
...		

* *isolated* means there were no edits in the given segment other than for this flag type

Table 1: Data analysis: Edit count by flag type

3.2 Common depassivization patterns

Let us digress for a moment from discussion of our experimental methodology to look at common depassivization patterns, as this will provide context to our analysis of the behaviour of our translation systems in the following subsection.

Based on inspection of a sampling of the edits performed by the human editor in response to *Avoid passives* flags, there is one canonical pattern for depassivizing German sentences and a second pattern, less common but still occurring repeatedly, that we refer to as the *verb-swap* pattern.

In a **canonical depassivization**, as illustrated in figure 1 (in English for the convenience of non-German-speaking readers),

- the subject noun phrase becomes an object

NP1-subj is V-PP → NP2-subj V-finite NP1-obj
e.g. 'The apple is eaten.' → 'The man eats the apple.'

Figure 1: Canonical depassivization pattern

noun phrase (NP1-subj becomes NP1-obj),

- the verb is changed from passive to active voice, which typically involves dropping the auxiliary verb (*is* or a related form) and changing the full verb from participle form to finite (V-PP becomes V-finite), and
- a new subject noun phrase (NP2-subj) is introduced.

The third point is problematic for automatic depassivization. The new subject typically does not appear in the original sentence (except sometimes in a prepositional phrase); deciding its identity requires context and world knowledge. The best we can reasonably hope for from an automatic system (that operates at surface level) is to insert a 'dummy' subject: 'X eats the apple'. By operating at the POS level, we bypass this issue.

Also, at the POS level, the first point is not necessarily detectable, in which case the pattern appears to consist of only two parts: the transformation of the verb and the introduction of a noun phrase.

NP1-subj is V1-PP → NP1-subj V2-finite
e.g. 'The image is shown.' → 'The image appears.'

Figure 2: Verb-swap depassivization pattern

In the **verb-swap pattern** (figure 2), the transitive verb in passive voice is replaced by a semantically related intransitive verb in active voice.

The decision of when to use this pattern and the choice of which verb to introduce are both lexical – they depend semantically on the original verb. Working at the POS level thus simultaneously complicates matters, by removing information required for deciding whether to use this pattern, and simplifies them, by freeing the system from having to select a specific replacement verb.

3.3 Preliminary Results

Using the data described in section 3.1, we train two Moses systems: one standard phrase-based and one tree-based.

Since passives in German often involve long-distance dependencies, tree-based SMT is intuitively more promising for this task.

Table 2 gives the BLEU scores (Papineni et al, 2002) achieved by the respective systems on our test set. They suggest that the tree-based system is indeed slightly better.

System	BLEU
standard	72.57
tree-based	73.56

Table 2: BLEU scores achieved by our two systems

However, manual analysis reveals that the BLEU score difference is misleading and that both result sets are equally bad. The problem is that the system applies parts of the depassivization patterns independently of each other, and independently of whether there is a passive in a given clause.

In figure 3, for example, we see an illustration of test item #3, in which the standard phrase-based system performs only half of the canonical pattern: it correctly translates the infinitival passive verb form VVPP VAINF to an active infinitive verb VVINF, but it fails to insert the missing subject. (We reiterate that the system input, output, and reference are the POS sequences; surface forms are shown for the reader's convenience.) The tree-based system produces the exact same result for this item.

The input string of test item #4 (figure 4), meanwhile, consists of two clauses, only the second of which contains a passive. The first clause should thus not be modified, but both systems (which, again, produce the exact same output string) perform half of the canonical depassivization, inserting a noun phrase. On the second clause, the systems perfectly perform the canonical depassivization, but, as we see in figure 4, the standard phrase-based system appears to be performing it as two independent changes. One change is the deletion of the verb participle VVPP and the other consists of the replacement of the finite auxiliary verb VAFIN with the finite full verb VVFIN and the insertion of a noun phrase. The tree-based system similarly performs these as two independent changes.

Both systems successfully apply the verb-swap pattern to item #1 (figure 5), producing output iden-

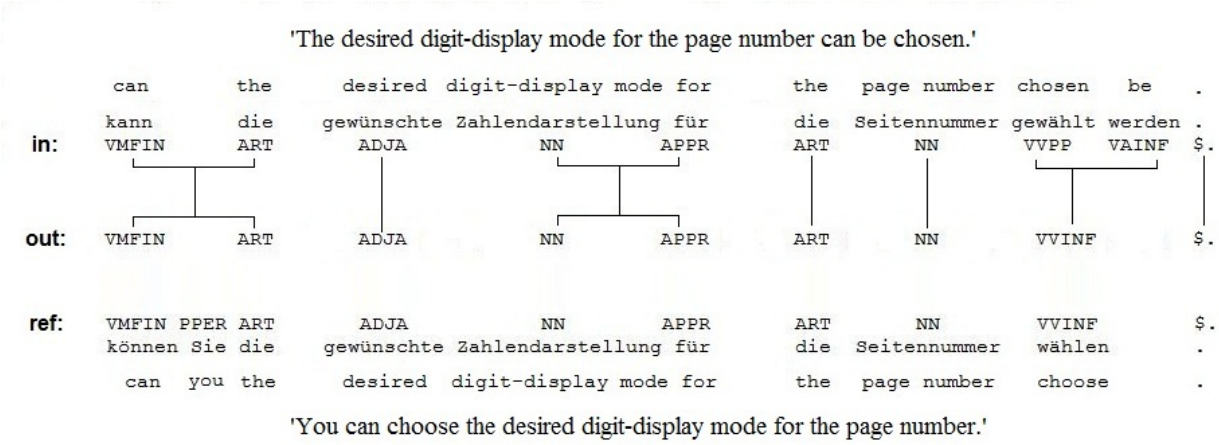

Figure 3: Test item #3, with output from the standard phrase-based system, including phrase correspondence

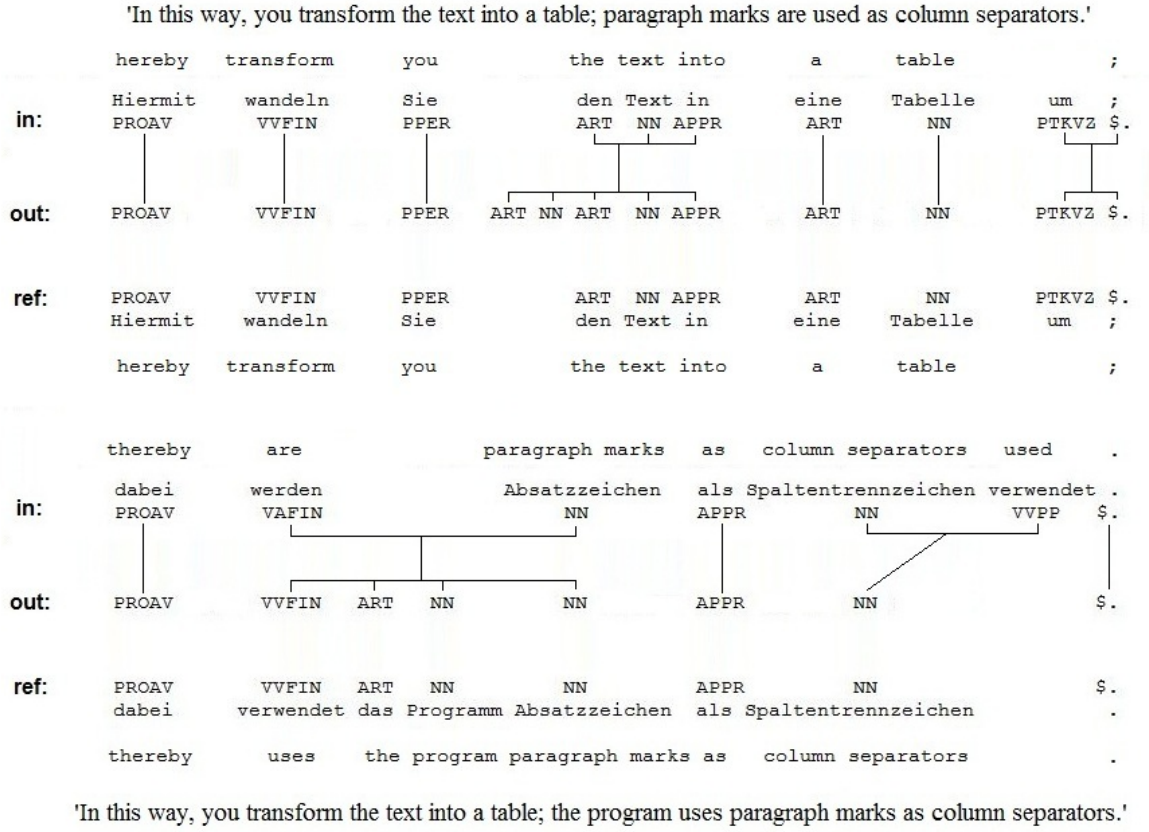

Figure 4: Test item #4, split into two lines due to its length, with output from the standard phrase-based system, including phrase correspondence

tical to the reference. However, the deletion of the verb participle VVPP (*gestartet* 'started') and the insertion of the semantically related finite verb VVFIN (*beginnt* 'begins') occur in different top-level phrases, which leads one to suspect that this success at the POS level would not easily be carried over to a success at the surface level.

4 Our Plan

To avoid the problem of the system performing depassivization steps where there was no passive to begin with, one could try giving the system information on where the *Avoid passives* flag occurred within the sentence. One way would be to treat a token within the flagged region as being of an entirely different class, e.g. a flagged finite auxiliary verb might be VAFIN_flagged rather than VAFIN.

Another idea for helping the system to learn where not to apply depassivization would be to add training data in which no passive occurs, and thus in which the source and target segments are identical.

The depassivization of German sentences often involves long-distance relationships on the input side which disappear in the output due to the elimination of the auxiliary verb. Braune et al (2012) extend hierarchical SMT with a method to extract an additional and separate set of rules specifically for long-distance reorderings. An SMT depassivizer such as ours may benefit from incorporation of their method. It seems therefore promising to investigate this in future work.

The 517 segment pairs containing depassivization form an excruciatingly small training set by the standards of SMT, so an obvious approach to improving the results is to get more data, which means collecting German passive sentences and depassivizing them by hand.

If an SMT system proves able to depassivize at the POS level, that would give us reason to expect that it could do the same at the surface level given enough data. That said, a POS-level SMT depassivizer could in itself perhaps be useful as a component of an automatic surface-depassivizer that uses heuristics to guess the output words from the alignments between the input words and the output POSes.

5 Conclusions

Using statistical machine translation technology, we produced systems that are sometimes able to depassivize German sentences represented at the part-of-speech level, though not with sufficient consistency to be useful. Nonetheless, our preliminary results show some possibility that this strategy has the potential to be successful. Our results are preliminary since we used a tiny data set consisting of 517, 20, and 10 text segment pairs for training, tuning, and test sets, respectively. We were limited to this size because the data is slow and expensive to produce, as each segment must be edited by a human. We presented ideas for improving the system, and if these prove fruitful and we are able to achieve an automatic German depassivizer, it opens the door to possibly automating the correction of a variety of other grammar and style issues in various languages using the same technique.

References

Fabienne Braune, Anita Gojun, and Alexander Fraser. 2012. Long-distance reordering during search for hierarchical phrase-based SMT. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pp. 177-184, Trento, Italy.

Andrew Bredenkamp, Berthold Crysmann, and Mirela Petrea. 2000. Building Multilingual Controlled Language Performance Checkers. In *Proceedings of the 3rd International Workshop on Controlled Language Applications*, pp. 83-89, Seattle, Washington.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. In *Proceedings of the Second Workshop On Statistical Machine Translation*, Prague, Czech Republic.

Dmitriy Genzel. 2010. Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pp. 376-384, Beijing, China.

Philipp Koehn and Hieu Hoang. June, 2007. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 868-876, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi,

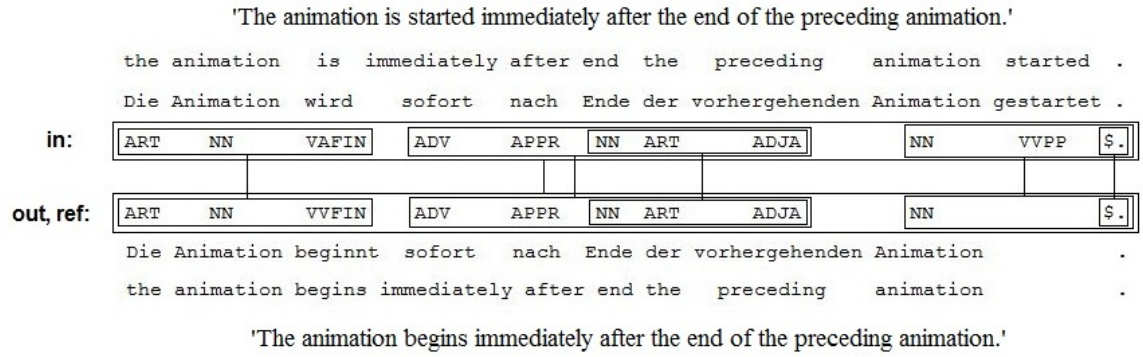

Figure 5: Test item #1, with output from the tree-based system, including phrase correspondence

Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. June, 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. July, 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, Pennsylvania.

Maja Popović and Hermann Ney. May, 2006. POS-based Word Reorderings for Statistical Machine Translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 1278-1283, Genoa, Italy.

tekom. 2011. *Regelbasiertes Schreiben: Deutsch für die Technische Kommunikation.*

Jörg Tiedemann. May, 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.