UdS-Sant: English–German Hybrid Machine Translation System

Santanu Pal1, Sudip Kumar Naskar2, Josef van Genabith1
1Universität des Saarlandes, Saarbrücken, Germany
2Jadavpur University, Kolkata, India
{santanu.pal, josef.vangenabith}@uni-saarland.de
sudip.naskar@cse.jdvu.ac.in

Abstract
This paper describes the UdS-Sant English–German Hybrid Machine Translation (MT) system submitted to the Translation Task organized in the Workshop on Statistical Machine Translation (WMT) 2015. Our proposed hybrid system brings improvements over the baseline system by incorporating additional knowledge such as extracted bilingual named entities and bilingual phrase pairs induced from example-based methods. The reported final submission is the result of a hybrid system obtained from confusion network based system combination that combines the best performance of each individual system in a multi-engine pipeline.

1 Introduction
In this paper, we present Universität des Saarlandes (UdS) submission (named UdS-Sant) to WMT 2015 using a Hybrid MT framework. We participated in the generic translation shared task for the English-German (EN-DE) language pair.

Corpus-based MT (CBMT) has delivered progressively improved quality translations since its inception. There are two main approaches to corpus-based MT – Example Based Machine Translation (EBMT) (Carl and Way, 2003) and Statistical Machine Translation (SMT) (Brown et al., 1993; Koehn, 2010). Out of these two, in terms of large-scale evaluations, SMT is the most successful MT paradigm. However, each approach has its own advantages and disadvantages along with its own methods of applying and acquiring translation knowledge from the bilingual parallel training data. EBMT phrases tend to be more linguistically motivated compared to SMT phrases which essentially operate on n-grams. The knowledge extraction as well as representation process, in both EBMT and SMT, uses very different techniques in order to extract resources. Even though, SMT is the most popular MT paradigm, it sometimes fails to deliver sufficient quality in translation output for some languages, since each language has its own difficulties.

Multiword Expressions (MWEs) and Named Entities (NEs) offer challenges within a language. MWEs are defined as idiosyncratic interpretations that cross word boundaries (Sag et al., 2002). Named entities on the other hand often consist of more than one word, so that they can be considered as a specific type of MWEs such as noun compounds (Jackendoff, 1997). Traditional approaches to word alignment such as IBM Models (Brown et al., 1993) are unable to tackle NEs and MWEs properly due to their inability to handle many-to-many alignments. In another well-known word alignment approach, Hidden Markov Model (HMM: (Vogel et al., 1996)), the alignment probabilities depend on the alignment position of the previous word. It does not explicitly consider many-to-many alignment either.

We address this alignment problem indirectly. The objective of the present work is threefold. Firstly, we would like to determine how treatment of MWEs as a single unit affects the overall MT quality (Pal et al., 2010; Pal et al., 2011). Secondly, whether a prior automatic NE aligned parallel corpus as well as example based parallel phrases can bring about any further improvement on top of that. And finally, whether system combination can provide any additional advantage in terms of translation quality and performance.

The remainder of the paper is organised as follows. Section 2 details the components of our system, in particular named entity extraction, translation memory, and EBMT, followed by description of 3 types of Hybrid systems and the system combination module. In Section 3, we outline the complete experimental setup for the shared task and
provide results and analysis on the performance on the test set in Section 4. Section 5 concludes the proposed research.

2 System Description

Our system is designed with three basic components: (i) preprocessing, (ii) hybrid systems and (iii) system combination.

2.1 Preprocessing

Data pre-processing plays a very crucial part in any data-driven approach. We carried out preprocessing in two steps:

- Cleaning and clustering sentences based on sentence length.
- Effective preprocessing of data in the form of explicit alignment of bilingual terminology (viz. NEs and MWEs).

The preprocessing has been shown (cf. Section 2.1.2) to improve the output quality of the baseline PB-SMT system (Pal et al., 2013; Tan and Pal, 2014).

2.1.1 Corpus cleaning

We utilized all the parallel training data provided by the WMT 2015 shared task organizers for English–German translation. The training data include Europarl, News Commentary and Common Crawl. The provided corpus is noisy and contains some non-German as well as non-English words and sentences. Therefore, we applied a Language Identifier (Shuyo, 2010) on both bilingual English–German parallel data and monolingual German corpora. We discarded those parallel sentences from the bilingual training data which were detected as belonging to some different language by the language identifier. The same method was also applied to the monolingual data.

Successively, the corpus cleaning process was carried out first by calculating the global mean ratio of the number of characters in a source sentence to that in a target sentence and then filtering out sentence pairs that exceed or fall below 20% of the global ratio (Tan and Pal, 2014). We sorted the entire parallel training corpus based on their sentence length. Tokenisation and punctuation normalisation were performed using Moses scripts. In the final step of cleaning, we filtered the parallel training data on maximum allowable sentence length of 100 and sentence length ratio of 1:2 (either direction). Approximately 36% sentences were removed from the total training data during the cleaning process.

2.1.2 Explicit Preprocessing of Terminologies

Two kinds of terminologies, viz. NEs and MWEs, were considered in the present work. Intuitively, MWEs should be both aligned in the parallel corpus and translated as a whole. However, state-of-the-art PB-SMT (or any other approaches to SMT) does not generally treat MWEs as special tokens. This is the motivation behind considering MWEs for special treatment in this work. By converting the MWEs into single tokens, we make sure that PB-SMT also treats them as a whole.

NE Alignment (NEA): For NE alignment, we first identify NEs on both sides of the parallel corpus using Stanford NER\(^1\). Next, we try to align the extracted source and target NEs. If both sides contain only one NE then the alignment is trivial, and we add such NE pairs to seed another parallel NE corpus that contains examples having only one token in both sides. Otherwise, we establish alignments between the source and target NEs using minimum edit distance method. For language pairs having different orthographies (e.g. English–Hindi) NE alignments can be established through transliteration (Pal et al., 2010). If both the source and target sides contain \(n\) number of NEs, and the alignments of \(n - 1\) NEs can be established through minimum edit distance method or by means of already existing alignments, then the \(n\)th alignment is trivial. The bilingual NE pairs extracted thus serve as additional training material and they improve the word alignment at the start of the MT pipeline.

MWE Identification: Translation correspondences between English MWEs and German MWEs are mainly many-to-one correspondences. Therefore, instead of extracting a bilingual MWE list between source and target, we identify the MWEs from the English training sentences and prepare an English MWE list. Once the MWEs are identified, they are converted into single tokens by replacing the spaces with underscores (“_”) so that their alignments can be mapped to single tokens. Before decoding, MWEs in the source side of the testset are also single tokenized by looking up the extracted MWE list. In this experiment, we have followed Point-wise Mutual Information.

\(^1\)http://nlp.stanford.edu/software/CRF-NER.shtml
mation (PMI), Log-likelihood Ratio (LLR), Phi-
coefficient and Co-occurrence measures for iden-
tification of MWEs on the English side. Finally,
a system combination model has been developed
which provides a normalized score for each of
the extracted MWEs. A predefined cut-off score
has been considered and the candidates having
scores above the threshold value are considered as
MWEs.

Example Based Phrase Extraction: We use
EBMT techniques to extract additional phrase
pairs from the training data to augment the SMT
(baseline) phrase pairs in our experiments. We ex-
tract EBMT phrase pairs based on the work de-
scribed in (Cicekli and GÜvenir, 2001), a compiled
approach of EBMT to automatically extract
translation templates from sentence-aligned bi-
ingual text. They observed the similarities and dif-
fences between two example pairs. Two types of
translation templates, i.e. generalized and atomic
templates, are extracted by applying this approach.
A generalized translation template replaces sim-
ilar or differing sequences with variables while
an atomic translation template does not contain
any variable. The atomic translation templates are
used as additional phrase pairs for our Hybrid MT
system. This particular approach has a cubic run-
time complexity with respect to the number of sen-
tences in the parallel corpus. It takes a significant
amount of time to extract phrase pairs even from a
small corpus. Therefore we used heuristics to re-
duce the time complexity. We divided the entire
corpus into \( n \) clusters based on sentence length
such that similar length sentences belong to the
same cluster. We extract atomic translations from
each of these clusters. For this task, we applied
EBMT phrases as addition parallel training ex-
template to explicitly enhanced the word alignment
model of the MT pipeline.

2.2 Hybrid System

The Hybrid approach is investigated by combining
multiple knowledge sources such as NEA, EBMT
Phrases and MWEs and followed different strate-
gies. As mentioned earlier, we implemented sev-
eral different systems, namely:

(1) Baseline PB-SMT,

(2) Baseline PB-SMT with NE alignment
    (NEA),

(3) NEA with EBMT phrase extraction (NEA–
    EBMT),

(4) NEA with EBMT phrase extraction and
    single-tokenised MWE (NEA–EBMT–
    MWE) and

(5) LM–NEA–EBMT–MWE hybrid system
    (see Section 2.2.1).

The baseline SMT system is trained on the cleaned
English-German parallel corpus. The NEA sys-
tem makes use of NE aligned parallel data as ad-
tional parallel examples. Similarly, EBMT phrase
pairs as well as NE aligned data are also used as
additional training example in the NEA–EBMT
system. The NEA–EBMT–MWE system is very
similar to the above mentioned the NEA–EBMT
system, the only difference being that the identi-
fied source side English MWEs are converted into
single tokens for NEA–EBMT–MWE. In order to
achieve optimal performance from the component
modules, we finally generated a composite transla-
tion output using confusion network-based system
combination (cf. Section 2.3).

2.2.1 LM-NEA-EBMT-SMT hybrid system

In this system, we experiment with the above de-
scribed models with varying size of monolingual
data. We experimented with 4 folds of monolin-
gual data to train the language Models (LM):

- \( \text{LM}_1 \): Only using the target side (i.e. Ger-
    man) of the parallel training data \((L)\) for lan-
guage modeling
- \( \text{LM}_2 \): \( L + \) double size of \( L \) in terms of num-
    ber of sentences, collected from the cleaned
    monolingual corpus
- \( \text{LM}_3 \): \( L + \) triple size of \( L \) from the cleaned
    monolingual corpus
- \( \text{LM}_4 \): \( L + \) all the cleaned monolingual data

Therefore, finally there were 16 different sys-
tems (4 systems, i.e., Baseline, NEA, NEA–
EBMT and NEA–EBMT–MWE, each with 4 LM
settings) output available for system combination.

2.2.2 Post-processing

As a final step, we try to generate translations of
out-of-vocabulary (OOV) words that remain un-
translated in the output. These OOV words may
include some NEs that are already there in the parallel NE list, however they might remain untranslated during decoding. Our system post processed the output by replacing each such OOV NE with the corresponding target language NE after looking up the extracted NE list from the parallel corpus (cf. Section 2.1.2).

2.3 System Combination

System Combination is a technique, which combines translation hypotheses (outputs) produced by multiple MT systems. We applied a system combination method on the outputs of the different MT system described earlier. We implement the Minimum Bayes Risk coupled with Confusion Network (MBR-CN) framework described in (Du et al., 2009). The MBR decoder (Kumar and Byrne, 2004) selects the single best hypothesis from amongst the multiple candidate translations by minimising BLEU (Papineni et al., 2002) loss. This single best hypothesis serves as the backbone (also referred to as skeleton) of the confusion network and determines the general word order of the confusion network. A confusion network (Matusov et al., 2006) is built from the backbone while the remaining hypotheses are aligned against the backbone using METEOR (Lavie and Agarwal, 2007) and the TER metric (Snover et al., 2006). The features used to score each arc in the confusion network are word posterior probability, target language model (3-gram, 4-gram), and length penalties. Minimum Error Rate Training (MERT) (Och, 2003) is applied to tune the CN weights (Pal et al., 2014).

3 Experiment Setup

3.1 Baseline Settings

The effectiveness of the present work is demonstrated by using the standard log-linear PB-SMT model as our baseline system. For building the baseline system, we used a maximum phrase length of 7 and a 5-gram language model. The other experimental settings were: SymGIZA++ aligner (Junczys-Dowmunt and Szal, 2012), which is a modified version of GIZA++ word alignment models by updating the symmetrizing models between chosen iterations of the original word alignment training algorithms and phrase-extraction (Koehn et al., 2003). The reordering model was trained on hier-mslr-bidirectional (i.e. using both forward and backward models) and conditioned on both source and target language. The reordering model was built by calculating the probabilities of the phrase pairs being associated with the given orientation such as monotone (m), swap (s) and discontinuous (d). The 5-gram target language model was trained using KENLM (Heafield, 2011). Parameter tuning was carried out using both k-best MIRA (Cherry and Foster, 2012) and Minimum Error Rate Training (MERT) (Och, 2003) on a held-out development set. After the parameters were tuned, decoding was carried out on the held out testset.

Note that all the systems described in Section 2 employ the same PB-SMT settings (apart from the feature weights which are obtained via MERT) as the Baseline system.

4 Results and Analysis

As described in Section 2.2.1, we developed 16 different systems. Instead of using all these 16 different systems, we apply only the 6 best performing systems for system combination. Performance is measured on the devset. Table 1 reports the final evaluation results obtained on the test dataset. The best 6 systems are as follows:

- System 1: NEA–EBMT (selective high frequency phrases) with baseline PB-SMT settings and LM$_1$.
- System 2: System 1 experimental settings + single tokenised source MWEs (i.e. NEA–EBMT–MWE, cf. Section 2.2).
- System 3: System 2 with MIRA-MERT coupled tuning.
- System 4: System 3 with LM$_2$.
- System 5: System 3 with LM$_3$.

System 6 provides the individual best system. System combination (System-7 in Table 1) of the 6 best performing individual systems brings considerable improvements over each of the individual component systems.

5 Conclusions and Future Work

A hybrid system (System 6) with NE alignment, EBMT phrases, single-tokenized source MWEs, and MIRA-MERT coupled tuning results in the best performing system. However, confusion
network-based system combination outperforms all the individual MT systems. The fact that the systems were tuned with BLEU scores may be one of the reasons behind the poor TER scores produced by the systems. In future, we will carry out in depth investigation of the impacts of MWEs within the current experimental settings. We will also analyze the usability and contribution of the novel EBMT phrases in the SMT decoder.

Acknowledgments

The research leading to these results has received funding from the EU FP7 Project EXPERT - the People Programme (Marie Curie Actions) (Grant No. 317471)

References


