Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Machine Translation with Significant Word Reordering and Rich Target-Side Morphology

Bushra Jawaid

Ph.D. Thesis Proposal

Draft: January 7, 2013

Supervisor: RNDr. Ondřej Bojar, Ph.D.
Institute of Formal and Applied Linguistics (ÚFAL)
Faculty of Mathematics and Physics
Charles University in Prague
Malostranské náměstí 25, 118 00 Praha 1
Czech Republic

Thesis Committee: RNDr. Zedněk Žabokrtský (chair), ÚFAL MFF UK
prof. RNDr. Jan Hajič (vice-chair), ÚFAL MFF UK
prof. PhDr. Eva Hajičová, ÚFAL MFF UK
RNDr. Vladislav Kubon, ÚFAL MFF UK
RNDr. Kiril Ribarov, České energetické závody a.s.
RNDr. Jan Cuřín, IBM ČR

Contents

Chapter 1

Introduction

1.1 Problem Statement

The aim of the thesis is to explore methods of machine translation when translating between languages with significant differences in word order and in the direction to the morphologically richer language, such as English-to-Urdu translation. Both these aspects increase the complexity of the translation task, strengthening the data-acquisition bottleneck and enlarging the search space of possible hypotheses. The main motivation for the thesis is thus to model the problem of reordering and word form choice separately to achieve an improved generalization and perform well on unseen sentences.

1.2 Proposed Solution

Statistical machine translation (SMT) systems and in particular phrase-based SMT systems (Section 2.2.1) usually don't perform well for the language pairs that differ in sentence structure (Koehn et al., 2009) and when target language is rich in inflection. We thus focus on translating in the hard direction i.e. translating from morphologically poor to morphologically richer languages and with source and target languages differing in syntactic structure.

We propose the integration of morpho-syntactic information in phrase-based and syntax-based Machine Translation systems. We intend to use hierarchical or surface syntactic models for languages of large vocabulary size and improve the translation quality using two-step approach (Fraser, 2009; Fraser et al., 2012). The two-step scheme basically reduces the complexity of hypothesis construction and selection by separating the task of source-to-target reordering from the task of generating fully inflected target-side word forms. In the first step, reordering is performed on the source data to make it structurally similar to the target language and in the second step, lemmatized target words are mapped to fully inflected target words. In Chapter 3, we will first introduce the reader to the detailed architecture of the two-step translation setup and later its further proposed enhancements for dealing with the above mentioned issues.

To investigate these issues we have chosen English-Urdu and English-Czech language pairs. Our first language pair i.e. English-Urdu exhibits both of these characteristics which shows complexity of modeling this language pair for the translation task; English is SVO (subject-verb-object) language whereas Urdu follows SOV sentence structure which requires translation system to move verb to the end of the sentence when translating from English to Urdu. Both target languages i.e. Urdu as well as Czech are also morphologically rich. For instance, adjectives in Urdu are inflected according to the gender and number of the following noun and verbs take gender of either subject or object depending on the tense and aspect. The morphological richness increases data sparseness and the differences in word order compel phrase-based SMT systems to learn long distance reordering. The author's Master thesis (Jawaid, 2010) already discussed translation issues in the direction from English to Urdu and proposed solutions to deal with the word order differences in the given language pair. This work is further extension of the author's previous research.

Chapter 2

Related Work

In this chapter we give the brief overview of statistical MT and formally introduce SMT systems that we use in our experiments. We further discuss significant research that has been done in integrating linguistic information to the SMT systems for improving the quality of translation.

2.1 Machine Translation

The task of a machine translation (MT) system is to translate the text from one language into text into another. The translation of source text into target requires deep understanding of linguistic knowledge at certain levels such as: morphology (dealing with rich inflectional nature of the languages), syntax (syntactical complexities such as difference in word order of source and target), semantics etc. Different approaches of MT deal with complex issues of the languages at various levels with differences in translation modeling techniques. MT is roughly classified into rule-based and data-driven paradigms. In classical rule-based systems, linguists perform deep analysis of linguistic phenomena of the given language pair and capture them in hand-written transformation rules which is a very labor-intensive task. The rules are later applied by an MT engine. On the other hand, data-driven approaches use large text corpora to automatically learn translation equivalences based on the real examples that are extracted from the corpus without adding any linguistic knowledge. Modern statistical machine translation (SMT) (Koehn, 2011) systems are the most prominent paradigm in machine translation that sligthly differs from the contemporary data-driven approaches by extracting the knowledge from large parallel corpora with added linguistic information.

2.2 Statistical Machine Translation (SMT)

2.2.1 Phrase-based Translation Model (PBTM)

Phrase-based translation model operates on sequences of words called phrases. It is based on the noisy channel model (Brown et al., 1990) approach which is well defined over Bayes decision rule. Bayes formula takes into consideration the language model probability and the probability of translating source phrase into best matching target phrase to obtain best output translation.

In phrase-based model source sentences are segmented into a number of phrases where each phrase gets translated into a target phrase. Target phrases might get reordered based on word order differences between source and target language. Moses (Koehn et al., 2007) is one of the statistical phrase-based MT system that by default uses *distance* reordering that allows movement of input phrases relative to previous phrase. The phrase movement over large distance means more expensive translation and it is thus seldom used.

Moses automatically learns from the parallel corpus of any given language pair. It also combines the language model capabilities for producing fluent output translation. Moses offers two types of translation models: phrase-based and tree-based (Section 2.2.3). We are experimenting with both translation models of Moses in this work (see Chapter 3).

2.2.2 Factored-based Translation Model (FBTM)

Factored-based translation model (Koehn and Hoang, 2007) integrates linguistic information on top of the non-linguistically motivated PBTM. Factor-based models treat each token as vector of factors as oppose to PBTM which is restricted to words only. They help in providing morphological, syntactical and any other linguistic information, in form of factors, directly to the translation model for direct modeling of these aspects.

In factored-based models decoding process is splitted into mapping and generation steps. In mapping step, source factors are translated to target factors whereas generation step generate additional output factors from already translated output factor during mapping step.

Implementation of factored translation models is available in Moses toolkit.

2.2.3 Syntax-based Machine Translation Systems

Factored-based models help in predicting syntatic order of target language by using part-of-speech tagged language model. Syntactic language models help translation model in deciding which translation options are more likely by looking at the sequence of output factor consist of part-of-speech. Beside factored models, prediction of accurate syntactic order of target language has been exploited in many different ways.

- Pre-reordering source text to match the word order of target language.
- Building syntactically motivated translation models.

Different approaches have been adapted for applying syntactic knowledge to the corpus before passing it to the translation system. For instance syntactic pre-reordering, syntactic reranking (post-processing) and many others. Syntactic pre-reordering has been shown effective many times for introducing syntax in SMT. So far syntactic pre-processing is applied on a source language in two different ways, either by using hand-crafted transformation rules or by learning transformation rules automatically from bitext. Jawaid and Zeman (2011) transformation system is based on the former approach, this approach was previously successfully applied to other language pairs (Collins et al., 2005; Wang et al., 2007; Ramanathan et al., 2008) as well. Later approach is adapted by many others: Nguyen and Shimazu (2006) introduced idea of using probabilites in reordering process and trained Baye's oriented transformation model using bitext and source language parser, Xia and McCord (2004) extracted both lexicalized and unlexicalized CFG rules from bitext using both source and target language parses and, Habash (2007) reordering rules are learnt using marked syntactic dependencies on source side that are aligned with target words in bitext, reordering rules are then used to preprocess both training and test corpus.

Li et al. (2007) first gave idea of using maximum entropy model based on source language parse trees to get n-best syntactic reorderings of each sentence which was further extended to the use of lattices. After Niehues and Kolss (2009), Bisazza and Federico (2010) further explored lattice-based reordering techniques for Arabic-English; they used shallow syntax chunking of the source language to move clause-initial verbs up to the maximum of 6 chunks where each verb's placement is encoded as separate path in lattice and each path is associated with a feature weight used by the decoder.

Many contributions have been made in the direction towards syntactic knowledge-oriented translation models. Wu (1997); Yamada and Knight (2001) and many others proposed translation systems similar to Chiang (2005). Yamada and Knight (2001) used methods based on tree-to-string mappings where source language sentences are first parsed and later operations on each node such as reordering child nodes, inserting extra words at each node and translating leaf nodes are applied. In later research, Eisner (2003) presented issues of working with isomorphic trees and presented a new approach of non-isomorphic tree-to-tree mapping translation model using synchronous tree substitution grammar (STSG). Nevertheless, Melamed (2004) gave an idea of syntax-based SMT system using synchronous parsers.

Chiang (2005) proposed hierachical phrase-based system (Hiero) built on top of phrase-based sytem where simple phrases are replaced by heirarchical phrases. His model is formally a synchornous context-free grammar learned from bitext without using any explicit linguistic information. Joshua (Li et al., 2010) is another SMT system that uses *hierarchical phrase-based model* introduced by Chiang (2005). Joshua is also formally based on SCFG where rules are learnt from bitext during training.

Hoang and Koehn (2010) implemented tree-based translation model in Moses (Koehn et al., 2007), formally known as *hierarchical phrase-based model* and *syntax-based model*, sometimes also referred as *moses-chart*. In PBTM, translation process is carried out from left-to-right of input whereas TBTM builds translation options recursively. The main motivation of TBTM is to introduce syntax using tree structures and for that it uses Synchronous Context-Free Grammar (SCFG) as the underlying formalism. SCFG represents sentence-pairs of source and target languages as pairs of constituency trees. Grammar rules are automatically learned during training from bitext and they consist of both linguistically motivated non-terminals (NP, VP, ...; making the syntax-based model) as well as generic non-terminal (X; making the hierarchical model). The hierarchical model can be trained similarly to phrase-based models but the training of the syntax-based model requires syntactically annotated input.

Instead of using simple phrases, hierarchical model of Moses uses hierarchical phrases i.e. phrases that contain sub-phrases. In hierarchical model all grammar rules consist of only non-terminal (X) with the exception of two special *gluing rules* that uses S to combines sequences of X for generating final output.

Sentence translation probability is calculated using language model probability and the product of weights of all grammar rules use to construct the output translation. Weight of each grammar rule is calculated using log-linear model. Beside these main components of sentence generation, other scoring functions are also used.

Joshua is more or less equivalent to hierarchical model of Moses and translations are also scored in similar fashion as described for Moses TBTM.

Chapter 3

Two-Step Approach of Machine Translation

Factored translation models (Koehn and Hoang, 2007) come into play when one of the source or target language is morphologically rich. Each token in the factored model consists of number of factors representing the surface form, lemma, POS tag, so on. Translation options are constructed in a sequence of mapping steps. Because each translation option needs to be fully constructed before the actual search takes place, there is a high risk of combinatorial explosion of the search space (Bojar and Kos, 2010).

The linguistically motivated setups as used in factored-based models are prohibitively expensive for large data, see also Bojar et al. (2009). A number of researchers have thus tried diving the complexity of search into two independent phases: (1) translation and reordering, and (2) conjugation and declination i.e. for en-ur language pair, in the first step, source English gets translated to a simplified Urdu and in the second step, the simplified Urdu gets fully inflected. The most promising results were obtained with the second step predicting individual morphological features using a specialized tool (Toutanova et al., 2008; Fraser et al., 2012).

3.1 Basic Two-Step Setup

Our baseline system will be similar to the systems presented by Bojar and Kos (2010) and Fraser (2009). They used Moses in the first step which produces augmented simple target output. Output of the first step is not fully inflected target instead it represents *middle language* consist of lemma and other morphological features. The second step translation is monotone where another Moses system is trained on augmented lemmatized target input and fully inflected target output.

3.2 Proposed Configurations of 1st-Step Translation

In this section we provide the details of further refinement techniques for two-step baseline system that will model reordering in more elegant way instead of relying only on Moses default reordering system. We will also try to deal with the inflection prediction task cleverly.

3.2.1 Reordering Techniques

We plan to use more sophisticated systems for dealing with word reordering issues. We will replace phrase-based Moses on the first step with either Joshua or Moses-chart. These SMT systems allow block movements which could help in improving reordering. The output of the first step will consist of the series of *strings* representing 1-best reordering for each sentence.

To make the reordering task slightly easier for the first step's systems, we will first pre-reorder the input data and try to make the source and target word orders more similar to each other. For translating from English-to-Urdu, the data will be pre-reordered using the transformation system used in (Jawaid, 2010). The transformation system will produce 1-best reordered output which will be used as input for Joshua and Moses-chart.

3.2.2 Rule-based Pre-reordering

For overcoming the "hard decisions" that are encountered due to relying on one possible reordering of each sentence which cannot be undone during decoding phase, we will modify our transformation system to produce multiple reorderings of each sentence that will be later fed into the Moses. We leave the decision on Moses to pick the best reordering among several possible reorderings. Niehues and Kolss (2009) first used lattice-based pre-reordering approach where different possible reorderings of each sentence (collected by applying discontinuous non-deterministic POS rules learned from word-aligned corpus) encoded as weighted edges in lattice.

3.3 Exploring Intermediate Language

3.3.1 Lattices of N-best Hypothesis

In all the settings described above, the systems in the first step always produce the strings of 1-best reordered output that are later used by the second step. We further plan to extend the string-based output of the first step in the form of *word lattice* (Dyer et al., 2008) i.e. multiple reorderings of each input sentence will be produced, giving the second step systems the freedom to choose among reordered sentences the one that is the easiest to inflect.

3.3.2 Language of lofs and mots

In the middle language, we have planned to experiment with one or two factors. These factors are either joined together to create a single token or they are used as separate factors in Moses factored-based translation model. For presentation purpose we call them: "LOF" ("lemma or form", i.e. a representation of the lexical information) and "MOT" ("modified tag", i.e. representing the morphological properties). In the single-factor experiments the LOF and MOT are simply concatenated into a token in the shape LOF+MOT.

Figure 3.1 illustrates the range of LOFs and MOTs we have experimented with (Section 3.5.5). LOF_0 and MOT_0 are identical to the standard Czech lemma and morphological tag as used e.g. in the Prague Dependency Treebank (Hajič et al., 2006).

LOF_1 and MOT_1 together make what Bojar and Kos (2010) call "pluslemma". MOT_1 is less complex than the full tag by disregarding morphological attributes not generally overt in the English source side. For most words, LOF_1 is simply the lemma, but for frequent words, the full form is used. This includes punctuation, pronouns and the verbs "být" (to be) and "mít" (to have).

MOT_2 uses a more coarse grained part of speech (POS) than MOT_1. Depending on the POS, different attributes are included: gender and number for nouns, pronouns, adjectives and verbs; case for nouns, pronouns, adjectives and prepositions; negation for nouns and adjectives; tense and voice for verbs and finally grade for adjectives. The remaining grammatical categories are encoded using POS, number, grade and negation.

Word Form	LOF_0	LOF_1	MOT_0	MOT_1	MOT_2	Gloss
lidé	člověk	člověk	NNMP1-----A---1	NPA-	NMP1-A	people
by	být	by	Vc-------------	c---	V-----	would
neočekávali	očekávat	očekávat	VpMP---XR-NA---	pPN-	VMP-RA	expect

Figure 3.1: Examples of LOFs and MOTs used in baseline experiments.

LOFs and MOTs are examples of the possible configurations of the two-step scenarios to be experimented. In future, we expect to experiment with cleverer intermediate language options.

3.4 Proposed Configurations of 2nd-Step Translation

Phrase-based Moses in the second step will be replaced with the classifier discussed later in Section 3.4.1. The classifier takes a string in the middle language as input and outputs the fully inflected target words.

3.4.1 Maximum Entropy-Based Classifier

In this work, we plan to build a maximum-entropy-based classifier (McCallum et al., 2000) for inflection prediction task. The motivation behind introducing the classifier is to facilitate the use of features looking far away from the processed word. In the simple design of the two-step approach by Bojar and Kos (2010); Fraser (2009), the prediction was performed using a simple n-gram model so only few previous words helped in the decision. Perhaps a more important flaw of the simple design is that the few previous words, if not relevant, increase the sparsity and thus make the inflection decision harder.

Fraser et al. (2011) have tried two-step setup for English to German by replacing Moses at the second step with four HMMs (Hidden Markov Models). Each HMM is trained to predict the feature value of linguistic features such as gender, number, definiteness and case. Surface form is then generated by providing predicted feature values, POS tag and stem to morphological analyzer for German, SMOR.

Recently, Fraser et al. (2012) have used four linear chain CRFs (Conditional Random Fields) for English to German languagae pair to predict fully specified German representation at the sencond step. They trained each CRF for a specific linguistic feature that they want to predict. The common features functions are used in all models whereas context of only predicted linguistic feature is included.

Our classifier approach is very similar to Toutanova et al. (2008). In their setup, the MT system generates only stems in the first step and produce an n-best list which is further sorted and augmented with the fully inflected word forms by the inflection prediction model in the second step. On the other hand, our setup generates augmented lemmatized output in the first step and outputs lattices which encode generally more translation candidates than n-best list. Jeong et al. (2010) further extended work of Toutanova et al. (2008) by integrating their discriminative lexicon model directly into the search within their tree-to-string-based SMT system.

In Table 3.1, we provide a brief summary of relevant morphological features of our two target languages. The values of these features have to be predicted from the source or surrounding target-side context.

Features	Urdu	Czech	Both
POS categories	42	11 main or 67 detailed	
Gender		neuter, inanimate	masculine, feminine
Number		dual	singular, plural
Person			1,2,3
Tense			present, past, future
Aspect	subjunctive, continuous		perfective, imperfective
Case	ergative, oblique		nominative, accusative, dative, genitive, locative, vocative, instrumental
Grade			positive, comparative, superlative

Table 3.1: Identified Morphological Features for Urdu and Czech

3.5 Work So Far

3.5.1 Morphological Annotation for Urdu

Urdu is a low-resource language with respect to even the core processing tasks like POS tagging or morphological analysis. We have examined performance of existing taggers for Urdu and determined that they do not reach sufficent coverage and accuracy. We have thus improved part-of-speech tagging for Urdu by using existing tools and data for the language. First, we convert the output of the existing morphological tools to a common representation and more importantly, we unify the different tagsets. Second, we train and evaluate a new tagger on the available annotated data (in our unified tagset) and finally, we implement and evaluate a "voting" scheme that combines the outputs of all available taggers.

On an independent test set, our tagger outperforms the other tools by far. We gain some further improvement by implementing a voting strategy that allows us to consider not only our tagger but also include suggestions by the other tools. The final combined tagger reaches the accuracy of 87.98%.

In ongoing work, we are trying to refine the voting strategy, making it more context-dependent, by adding one more custom tagger trained to pick the best tag. Also, we are aware that the current ensemble of taggers is somewhat impractical: three taggers have to be run and the final answer is available only after their voting. We hplan to run this complex ensemble on a large monolingual corpus and use this data to train a single, standalone tagger. We also plan to release the standalone tagger. In order to conduct two-step experiments for en-ur language pair, we plan to add back the detailed morphological information we are now stripping off.

3.5.2 Experimental Setup

We have carried out baseline two-step experiments for en-cs language pairs. For baseline setup, we use the Moses toolkit (Koehn et al., 2007) and GIZA++ (Och and Ney, 2000). The texts were processed using the Treex platform (Popel and Žabokrtský, 2010)[1], which included lemmatization and tagging by Morce (Spoustová et al., 2007). After the tagging, we tokenized further so words like "23-year" or "Aktualne.cz" became three tokens.

In the first step, source English gets translated to a simplified Czech and in the second step, the simplified Czech gets fully inflected. On both steps we have used either phrase-based or factored-based models of Moses, depending on the construction of the middle language.

3.5.3 En-Cs Data

Our training data for en-cs experiments is summarized in Table 3.2. In all experiments reported here, we have used the Small dataset only. The language model (LM) for these experiments is a 5-gram one based on the target-side of Small only.

Dataset	Sents (cs/en)	Toks (cs/en)	Source
Small	197k parallel	4.2M/4.8M	CzEng 1.0 news
Large	14.8M parallel	205M/236M	CzEng 1.0 all
Mono	18M/50M	317M/1.265G	WMT12 mono

Table 3.2: Summary of en-cs training data.

3.5.4 Decoding Paths in Two-Step Setups

Each of the searches in the two-step setup can be as complex as the various single-step configurations (Bojar et al., 2012). We test just one decoding path for the one or two factors in the middle language.

All experiments with one middle factor (i.e. "+") follow this config: tF-LOF+MOT = tLOF+MOT-F, i.e. two direct translations where the first one produces the concatenated LOF and MOT tokens and the second one consumes them. The first step uses a 5-gram LOF+MOT language model and the second step uses a 5-gram LM based on forms.

This setup has the capacity to improve translation quality by producing forms of words never seen aligned with a given source form. For example the English word *green* would be needed in the parallel data with all the morphological variants of

[1] `http://ufal.mff.cuni.cz/treex/`

the Czech word *zelený*. Adding the middle step with appropriately reduced morphological information so that only features overt in the source are represented in the middle tokens (e.g. negation and number but not the case) allows the model to find the necessary form anywhere in the target-side data only:

$$\textit{green} \rightarrow \textit{zelený+NSA-} \rightarrow \begin{cases} \text{(genitive) } \textit{zeleného} \\ \text{(dative) } \textit{zelenému} \\ \ldots \end{cases}$$

The experiments with two middle factors (i.e. “|”) use this path: tF-LOFaMOT = tLOFaMOT-F:LOF-F. The first step is identical, except that now we use two separate LMs, one for LOFs and one for MOTs. The second step has two alternative decoding paths: (1) as before, producing the form from both the LOF and the MOT, and (2) ignoring the morphological features from the source altogether and using just target-side context to choose an appropriate form of the word. This setup is capable of sacrificing adequacy for a more fluent output.

3.5.5 Experiments with Two-Step Setups

Two-step setups can use factors in the source, middle or the target language. We have experimented with factors only in the middle language (affecting both the first and the second search) and use only the form in both source and target sides.

Table 3.3 reports the BLEU scores when changing the number of factors (“+” vs. “|”) in the middle language, the type of the LOF and MOT and also two options affecting reordering capacity of the second step: “-max-phrase-length” and “-distortion-limit”. We decided to experiment with these two options because some of our initial runs had accidentally reordering allowed in the second step and gained some improvement. The setting 1 and 0, respectively, is the standard monotone word-for-word translation in the second step. 10–6 are the default settings of Moses that allow reordering within phrases (up to 10 words away) and also reordering of whole phrases (up to 6 words away, subject to reordering penalty). The last setting, 10–0 corresponds to reordering only within phrases and no reordering of phrases. The main drawback is that our training data for the second step do not come from the real run of the first step (which would allow the second step to learn to post-edit first step outputs), but rather from GIZA word alignments, which is essentially just a noisy identity alignment.

Baseline setup for 2-step experiments is choosen to be the same that was earlier reported in (Bojar and Kos, 2010), detailed description of baseline setup is also described in that paper. In (Bojar and Kos, 2010), intermediate language is represented as single token i.e. lof and mot0 combined together using “+” sign to form single token. According to our initial hypothesis, lof+mot0 forms are themselves causing data sparseness, so we formalized our two-step experiments in two directions:

- Intermediate Czech will be represented as single token i.e. lof+mot.
- Or, intermediate Czech will be represented as multi-factored tokens i.e. lof|mot.

-max-phrase-length	1	10	10
-distortion-limit	0	6	0
$\text{LOF}_0 \mid \text{MOT}_0$	12.42±0.48	10.71±0.43	12.47±0.49
$\text{LOF}_1 \mid \text{MOT}_1$	11.85±0.42	11.98±0.47	11.97±0.48
$\text{LOF}_1 \mid \text{MOT}_2$	12.47±0.51	12.50±0.51	12.49±0.51
$\text{LOF}_0 + \text{MOT}_0$	11.11±0.48	11.79±0.51	11.84±0.51
$\text{LOF}_1 + \text{MOT}_1$	12.10±0.48	12.24±0.48	12.25±0.49
$\text{LOF}_1 + \text{MOT}_2$	11.87±0.51	11.90±0.50	11.91±0.50

Table 3.3: Two-step experiments.

Table 3.3 shows results of single and multi token experiments. We see an interesting difference between MOT_1 and $\text{MOT}_{2\text{ or }0}$. The more fine-grained $\text{MOT}_{2\text{ or }0}$ work better in the two-factor "|" setup that allows to disregard the MOT, while MOT_1 works better in the direct translation "+".

Overall, we see no improvement over the tF-LOF+MOT=tLOF+MOT-F baseline (BLEU of 11.84) and this is mainly due to to the fact that we used Small data in both steps.

Chapter 4

Summary and Timeline

4.1 Summary

We have presented several techniques to deal with data sparsity and word reordering issues. We are trying to reduce the complexity of the search space and the risk of search errors that are mostly encountered due to modeling both reordering and morphology at the same step. We plan to split the two problems into separate steps. Target-specific morphological features are introduced in the second step only whereas morphological features common to both source and target together with word reordering are handled in the first step. In the second step, all remaining morphological features of the target language are decided based on monolingual information only. This approach reduces the risk of the combinatorial explosion, because the target side of the first step is not cluttered with information not available and relevant for the source language and the transfer.

Although this is not the first time the two-step approach is presented, our work is still novel in terms of: the language pairs we are going to deal with and the integration of different reordering systems in the first step on top of the classifier.

4.2 Timeline

Work on this thesis should be completed by September 2014, with remaining individual tasks being carried out according to the timeline below.

- **January to March 2013**
  - Finish work in progress on Urdu morphology extraction.
  - Release Urdu stand-alone tagger.
- **April 2013**
  - Identification of morphological features to be used in intermediate language for ur and cs.

  - Baseline experimentation for en-ur language pair.

- **May to July 2013**
  - Getting working knowledge of maximum entropy-based classifier.
  - Implementation of the classifier.

- **August to September 2013**
  - Experimentation after replacing moses with classifier on 2nd step.

- **October to December 2013**
  - Experimentation with syntactic translation models on first step.
  - Experimentation with pre-reordering source using transformation system for en-ur language pair.
  - Modification of transformation system to produce multiple reordered sentences.

- **January to March 2014**
  - Experimentation with lattices in intermediate layer.
  - Experimentation with possible different configurations of 1st and 2nd step.

- **April to August 2014**
  - Finalization of the experimentation.
  - Thesis writing.

Bibliography

Arianna Bisazza and Marcello Federico. Chunk-based verb reordering in vso sentences for arabic-english statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 235–243, Stroudsburg, PA, USA, 2010. ISBN 978-1-932432-71-8.

Ondřej Bojar and Kamil Kos. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W10-1705`.

Ondřej Bojar, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, and Zdeněk Žabokrtský. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March 2009. Association for Computational Linguistics.

Ondřej Bojar, Bushra Jawaid, and Amir Kamran. Probes in a taxonomy of factored phrase-based models. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 253–260, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2393015.2393050`.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 16:79–85, June 1990. ISSN 0891-2017.

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the ACL '05*, pages 263–270, Stroudsburg, PA, USA, 2005. doi: http://dx.doi.org/10.3115/1219840.1219873.

Michael Collins, Philipp Koehn, and Ivona Kučerová. Clause restructuring for statistical machine translation. In *Proceedings of the ACL '05*, pages 531–540, Stroudsburg, PA, USA, 2005. doi: http://dx.doi.org/10.3115/1219840.1219906.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. Generalizing word lattice translation. In *Proceedings of the ACL*, pages 1012–1020, Columbus, Ohio, USA, June 2008.

Jason Eisner. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of the ACL '03 - Volume 2*, pages 205–208, Stroudsburg, PA, USA, 2003. ISBN 0-111-456789. doi: http://dx.doi.org/10.3115/1075178.1075217.

Alexander Fraser. Experiments in morphosyntactic processing for translating to and from german. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 115–119, Stroudsburg, PA, USA, 2009.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Fritzinger. Morphological generation of german for smt. In *Machine Translation and Morphologically-rich Languages. Research Workshop of the Israel Science Foundation*, Israel, January 2011. University of Haifa.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. Modeling Inflection and Word-Formation in SMT. In *Proc. of EACL 2012*. Association for Computational Linguistics, 2012.

Nizar Habash. Syntactic preprocessing for statistical machine translation. In *In Proceedings of the MT Summit XI.*, 2007.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4, 2006.

Hieu Hoang and Philipp Koehn. Improved translation with source syntax labels. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 409–417, Stroudsburg, PA, USA, 2010. ISBN 978-1-932432-71-8.

Bushra Jawaid. Statistical machine translation between languages with significant word order difference. In *Master's Thesis*, page 99. Univerzita Karlova v Praze & University of Malta, aug 2010.

Bushra Jawaid and Daniel Zeman. Word-order issues in english-to-urdu statistical machine translation. *The Prague Bulletin of Mathematical Linguistics*, (95):87–106, 2011. ISSN 0032-6585.

Minwoo Jeong, Kristina Toutanova, Hisami Suzuki, and Chris Quirk. A discriminative lexicon model for complex morphology. In *Ninth Conference of the Association for Machine Translation in the Americas*, 2010.

Philip Koehn, Alexandra Birch, and Ralf Steinberger. 462 machine translation systems for europe. In *Proceedings of Machine Translation Summit XII*, 2009.

Philipp Koehn. Statistical machine translation. Cambridge University Press, 2011. ISBN 978-1-932432-09-1.

Philipp Koehn and Hieu Hoang. Factored translation models. In *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL*, pages 868–876, 2007.

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. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, 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 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P07/P07-2045`.

Chi-ho Li, Dongdong Zhang, Mu Li, Ming Zhou, Minghui Li, and Yi Guan. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of ACL*, pages 720–727, 2007.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Ann Irvine, Sanjeev Khudanpur, Lane Schwartz, Wren N. G. Thornton, Ziyuan Wang, Jonathan Weese, and Omar F. Zaidan. Joshua 2.0: a toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies. In *Proceedings of the WMT '10*, pages 133–137, Stroudsburg, PA, USA, 2010. ISBN 978-1-932432-71-8.

Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 591–598, San Francisco, CA, USA, 2000. ISBN 1-55860-707-2.

I. Dan Melamed. Statistical machine translation by parsing. In *In Proceedings of Association for Computational Linguistics*, 2004.

Thai Phuong Nguyen and Akira Shimazu. Improving phrase-based statistical machine translation with morpho-syntactic analysis and transformation. In *In 5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts, 2006.

Jan Niehues and Muntsin Kolss. A pos-based model for long-range reorderings in smt. In *Proceedings of the StatMT '09*, pages 206–214, Stroudsburg, PA, USA, 2009.

Franz Josef Och and Hermann Ney. A Comparison of Alignment Models for Statistical Machine Translation. In *Proceedings of the 17th conference on Computational linguistics*, pages 1086–1090. Association for Computational Linguistics, 2000. ISBN 1-555-55555-1.

Martin Popel and Zdeněk Žabokrtský. TectoMT: Modular NLP Framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg, 2010. Iceland Centre for Language Technology (ICLT), Springer. ISBN 978-3-642-14769-2.

Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, Jayprasad Hegde, Ritesh Shah, M., and Sasikumar M. Simple syntactic and morphological processing can help english-hindi statistical machine translation. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*, 2008.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha, 2007.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P08/P08-1059`.

Chao Wang, Michael Collins, and Philip Koehn. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 737–745, 2007.

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.*, 23:377–403, September 1997. ISSN 0891-2017.

Fei Xia and Michael McCord. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1220355.1220428. URL `http://dx.doi.org/10.3115/1220355.1220428`.

Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of ACL '01*, pages 523–530, Stroudsburg, PA, USA, 2001.