WDS'11 Proceedings of Contributed Papers, Part I, 161–166, 2011. ISBN 978-80-7378-184-2 © MATFYZPRESS

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

B. Jawaid

Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. This paper describes the integration of morpho-syntactic information in phrase-based and syntax-based Machine Translation systems. We mainly focus on translating in the hard direction which is translating from morphologically poor to morphologically richer languages and also between language pairs that have significant word order differences. 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]. 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. 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. We plan to conduct experiments for two language pairs: English-Urdu and English-Czech.

1. Introduction

The task of a machine translation (MT) system is to translate the text from one language into text into another. MT approaches are roughly classified into rule-based and data-driven paradigms. In classical rule-based systems, linguists perform deep analyses 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. Modern statistical machine translation (SMT) [Koehn, 2010] systems extract the knowledge from large parallel corpora with added linguistic information.

SMT systems and in particular phrase-based SMT systems (PSMT) 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. Our first language pair i.e. English-Urdu exhibits the same language 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. Urdu and Czech are morphologically rich languages. For instance, adjectives in Urdu are inflected according to the gender and number of the following noun. The morphological richness increases data sparseness and the differences in word order compel PSMT 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.

The rest of this paper is organized as follows: In Section 2 we briefly describe the tools and resources we will use to build the two-step model. Then we describe the basic two-step architecture in Section 3 and our proposed improvement techniques to exploit morpho-syntactic information for SMT systems in Section 4. Later, we briefly introduce the recent contributions

from other researchers in improving SMT quality (Section 5). Finally, in Section 6 we provide a brief summary of our proposed translation scheme.

2. Relevant Tools

In this section we provide a short overview of all the available tools and resources that will be used at each step of our two-step setup. All of the details concerning the specific use of each tool are explained in Section 4.

2.1. Moses

Moses [Koehn et al., 2007] is a statistical phrase-based MT system that 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.

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 difference between source and target language. Moses 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.

Tree-based Translation Model (TBTM). Moses tree-based translation model [Hoang and Koehn, 2010] is 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.

2.2. Joshua

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

2.3. 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] and [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.

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.

A brief introduction to the proposed input features and target classification is provided in Section 4.3.

3. Two-Step 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 idea behind using two-step translation is to avoid the explosion of the search space by dealing with reordering and word inflections in 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. This 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.

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.

Recently, [Fraser et al., 2011] has tried two-step setup by replacing Moses at the second step with 4 HMMs (Hidden Markov Models).

4. Proposed Configurations of Two-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.

4.1. 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.

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, transformation system will produce multiple reorderings of each sentence that will be later fed into the Moses in the form of a *word lattice* [Dyer et al., 2008]. 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.

4.2. Exploring Middle Layer

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 to the lattice-based output 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.

4.3. Using Classifier

Phrase-based Moses in the second step will be replaced with the classifier previously introduced in Section 2.3. The classifier takes a string in the middle language as input and outputs the fully inflected target words. In Table 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.

Table 1. Identified Morphological Features for Urdu and Czech

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

5. Related Research

Significant research has been done in integrating linguistic information to the SMT systems including syntactically motivated translation models and introducing syntax in phrase-based SMT systems.

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).

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. Our transformation system [Jawaid and Zeman, 2011] 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.

[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 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.

6. Conclusion

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. In the first step, only reordering and morphological features common to both languages are handled. In the second step, all remaining morphological features of the target language are decided based on monolingual information only. 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 classifier.

Acknowledgments. The work on this project was supported by the grant LC536 Centrum komputační lingvistiky of the Czech Ministry of Education.

References

Bisazza, A. and Federico, M., 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, pp. 235–243, Stroudsburg, PA, USA, 2010.

Bojar, O. and Kos, K., 2010 failures in english-czech phrase-based mt, in *Proceedings of the WMT '10*, pp. 60–66, Stroudsburg, PA, USA, 2010.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S., A statistical approach to machine translation, *Comput. Linguist.*, *16*, 79–85, 1990.

Chiang, D., A hierarchical phrase-based model for statistical machine translation, in *Proceedings of the ACL '05*, pp. 263–270, Stroudsburg, PA, USA, 2005.

Collins, M., Koehn, P., and Kučerová, I., Clause restructuring for statistical machine translation, in *Proceedings of the ACL '05*, pp. 531–540, Stroudsburg, PA, USA, 2005.

Dyer, C., Muresan, S., and Resnik, P., Generalizing word lattice translation, in *Proceedings of the ACL*, pp. 1012–1020, Columbus, Ohio, USA, 2008.

Eisner, J., Learning non-isomorphic tree mappings for machine translation, in *Proceedings of the ACL '03 - Volume 2*, pp. 205–208, Stroudsburg, PA, USA, 2003.

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

Fraser, A., Weller, M., Cahill, A., and Fritzinger, F., Morphological generation of german for smt, in *Machine Translation and Morphologically-rich Languages. Research Workshop of the Israel Science Foundation*, University of Haifa, Israel, 2011.

Hoang, H. and Koehn, P., Improved translation with source syntax labels, in *Proceedings of the WMT '10*, pp. 409–417, Stroudsburg, PA, USA, 2010.

Jawaid, B., Statistical machine translation between languages with significant word order difference, in *Univerzita Karlova v Praze & University of Malta*, p. 99, Praha, Czechia, 2010.

Jawaid, B. and Zeman, D., Word-order issues in english-to-urdu statistical machine translation, *The Prague Bulletin of Mathematical Linguistics*, pp. 87–106, 2011.

Jeong, M., Toutanova, K., Suzuki, H., and Quirk, C., A discriminative lexicon model for complex morphology, in *Ninth Conference of the Association for Machine Translation in the Americas*, 2010.

Koehn, P., *Statistical Machine Translation*, Cambridge University Press, 2010.

Koehn, P. and Hoang, H., Factored translation models, in *Proceedings of the EMNLP-CoNLL*, pp. 868–876, 2007.

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

Koehn, P., Birch, A., and Steinberger, R., 462 machine translation systems for europe, in *Proceedings of Machine Translation Summit XII*, 2009.

Li, C.-h., Zhang, D., Li, M., Zhou, M., Li, M., and Guan, Y., A probabilistic approach to syntax-based reordering for statistical machine translation, in *Proceedings of ACL*, pp. 720–727, 2007.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Irvine, A., Khudanpur, S., Schwartz, L., Thornton, W. N. G., Wang, Z., Weese, J., and Zaidan, O. F., Joshua 2.0: a toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies, in *Proceedings of the WMT '10*, pp. 133–137, Stroudsburg, PA, USA, 2010.

McCallum, A., Freitag, D., and Pereira, F. C. N., Maximum entropy markov models for information extraction and segmentation, in *Proceedings of the ICML '00*, pp. 591–598, San Francisco, CA, USA, 2000.

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

Ramanathan, A., Bhattacharyya, P., Hegde, J., Shah, M., R., and M., S., Simple syntactic and morphological processing can help english-hindi statistical machine translation, in *IJCNLP*, 2008.

Toutanova, K., Suzuki, H., and Ruopp, A., Applying Morphology Generation Models to Machine Translation, in *Proceedings of ACL-08: HLT*, pp. 514–522, Columbus, Ohio, 2008.

Wang, C., Collins, M., and Koehn, P., Chinese syntactic reordering for statistical machine translation, in *EMNLP-CoNLL*, pp. 737–745, 2007.

Wu, D., Stochastic inversion transduction grammars and bilingual parsing of parallel corpora, *Comput. Linguist.*, *23*, 377–403, 1997.

Yamada, K. and Knight, K., A syntax-based statistical translation model, in *Proceedings of ACL '01*, pp. 523–530, Stroudsburg, PA, USA, 2001.