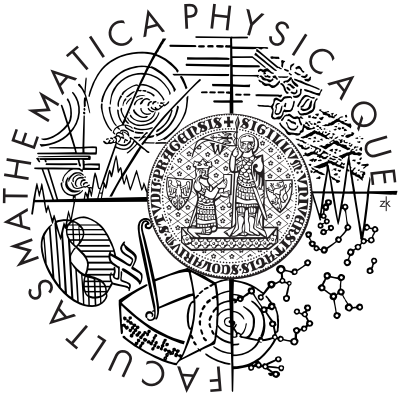
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



Hybrid Machine Translation

Amir Kamran

Ph.D. Thesis Proposal

Date: January 11, 2013

Supervisor: RNDr. Vladislav Kubon, 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. Ondřej Bojar, ÚFAL MFF UK
RNDr. Kiril Ribarov, České energetické závody a.s.
RNDr. Jan Cuřín, IBM ČR

Contents

1 Introduction

With the rise of social networking and the flood of information in foreign languages through web, the use of machine translation technology becomes inevitable. The current state-of-the-art systems are performing good for specific scenarios and languages. On one hand, the usefulness of the technology depends on the quality, on other hand, fully-automatic high-quality machine translation (FAHQMT) (Koehn, 2010) is still a cumbersome question.

The current approaches of machine translation are based on either symbolic/rule-based or statistical/corpus-based approaches (Habash et al., 2009). Both these approaches have their strengths and weaknesses. The symbolic approaches are built on linguistic knowledge and have the ability to deal with long distance dependencies, agreement and constituent reordering, hence they produce better structured translations. However, the deeper the linguistic analysis goes, the more expensive resources it will require. Resource richness are some of the main challenges for symbolic/rule-based approaches. Resource poverty often refers to the lack of monolingual and even multilingual resources involving the "poor" language (Habash et al., 2009).

On the other hand, the statistical approaches depends on the presence of parallel corpora and does not involve any deeper linguistic analysis. These approaches performs better in lexical selection and general fluency, as they are able to capture implicit knowledge contained in co-occurrence statistics. However, statistical systems usually generate structurally worse translations, since they model translation more locally and have problems with long distance reordering. For these approaches, the degree of resource poverty is defined primarily in terms of the size of monolingual and bilingual corpora for the language pair in question (Habash et al., 2009).

During the last few years, much interest has developed in the area of hybrid machine translation by researchers working on both symbolic and statistical approaches. This PhD proposal is focused on studying different ways of hybridization possible and applications of suitable hybrid machine translation approaches for the language pairs English-to-Urdu and English-to-Czech. The two main research ques-

tions in consideration are:

- The utilization of linguistic analysis in an optimal way to achieve better structural translation in the framework of statistical machine translation.
- How the hybrid machine translation can help overcome the resource poverty of a source or target language?

2 Related Work

A machine translation (MT) system based on combination of approaches is an interesting challenge. The main objective of hybridization is to take advantage of the strengths of both linguistic rules and statistical techniques. We can categorize the hybrid systems into two main groups **Single-Engine Hybridization** or **Multi-Engine Hybridization**.

2.1 Single-Engine Hybridization (SEH)

The main architecture of this scheme is based on a single approach either rule-based or statistical, but one or more steps of the MT pipeline are modified using the respective other approach. Thurmair (2009) refers this type of hybrid configuration as "Architecture Extensions". The modifications are possible at various levels depending on the targeted problem. For example, a statistical approach can be enhanced by reordering the source side of parallel corpus to handle the word order issues (Jawaid and Zeman, 2011).

We can further classify SEH into two categories based on the main approach used i.e. statistical machine translation (SMT) modified by linguistic information and rule-based machine translation (RBMT) modified by statistical techniques.

2.1.1 Modifications in SMT

Morphology based modifications Data sparseness is one of the major issues in SMT, mainly due to morphologically rich languages. Morphological preprocessing can be used to reduce the data sparsity (Goldwater and McClosky, 2005). Many studies examined the effects of various kinds of tokenization, lemmatization and part-of-speech (POS) tagging and showed a positive effect on statistical MT quality. Bojar and Prokopová (2006) reported improvement in statistical word alignment for Czech. Toutanova et al. (2008) used morphology generation models as postprocessing and reported improvements for Arabic and Russian.

Syntax based modifications More recently a number of statistical MT approaches have included syntactic information as part of the preprocessing phase or the decoding phase. The word order differences or the distance based reordering problems can be tackled by syntax based transformations. The idea is to apply transformation rules to the source language parse tree to make the order of the source sentence closer to the target sentence. The efficiency of these methods has been shown on various language pairs including: French to English (Xia and McCord, 2004), German to English (Collins et al., 2005), English to Chinese (Wang et al., 2007), English to Hindi (Ramanathan et al., 2008) and English to Urdu (Jawaid and Zeman, 2011). Quirk et al. (2005) used dependency trees and treelet phrase extraction to deal with word order differences between source and target languages.

Factor-based SMT Factored models (Koehn et al., 2007b) proposed a more integrated approach to enrich statistical MT with linguistic information. The basic idea behind factored translation models is to represent phrases not simply as sequences of fully inflected words, but instead as sequences containing multiple levels of information. A factored language model views a word as a vector of k factors:

$$w_i = f_i^1, f_i^2, ..., f_i^k$$

Factors can be anything, including morphological classes, stems, roots, and other such features in highly inflected languages (e.g., Arabic, German, Finnish, etc.), or data-driven word classes or semantic features useful for sparsely inflected languages (e.g., English). A two-factor language model is generated by standard class-based (Brown et al., 1992) language models, where one factor is the word class and the other is word itself. A factor-based model is a model over factors, i.e.

$$p(f_t^{1:k}|f_{t-1:t-n}^{1:k})$$

that can be factored as a product of probabilities of the form $p(f|f_1, f_2, ..., f_n)$. In factor-based statistical machine translation the task is twofold: One is to find an appropriate set of factors and two is to include an appropriate statistical model over those factors (Bilmes and Kirchhoff, 2003; Dep, 2008). Factored translation models closely follow the statistical modeling methods used in phrase-based models. Each of the mapping steps is modeled by a feature function. This function is learned from the training data, resulting in translation tables and generation tables (Koehn et al., 2007a). Koehn et al. (2007a) showed improvement for a number of language pairs using factor-based models. Ramanathan et al. (2009) applied factor-based models to English-Hindi language pair. Yeniterzi and Oflazer (2010) showed improvements with syntax-to-morphology mapping for factored translation models for English-to-Turkish. More recently Bojar et al. (2012) described a taxonomy of factored based scenarios for English to Czech and point out several common pitfalls when designing factored setups.

2.1.2 Modifications in RBMT

The direction where the RBMT system leads the translation and the SMT system provides complementary information, has been less explored. Data driven techniques have been exploited into different directions to improve the translation quality of RBMT systems:

- Preparing the resources for the RBMT systems e.g. by using automatically

extracted grammar rules from corpora. This type of modification termed as Pre-Editing by Thurmair (2009). Habash et al. (2009) pre-edit the RBMT system by enriching the dictionary with phrases from an SMT system.

- The other approaches are based on modifications in the core of the RBMT system by adding probabilistic information to the analysis and the parsing process. Attempts to improve the transfer selection process have also been made.

Pre-Editing implies modifications in language resources of rule-based systems such as dictionaries and grammar rules by using data driven techniques. Dictionaries are one of the major component of RBMT systems. By extracting terminologies from either monolingual or bilingual corpora, dicionary entries can be intelligently learnt. Monolingual corpora help in finding missing entries in the dictionaries whereas bilingual corpora help in finding translation candidates.

Although these approaches are already successfully applied in rule-based systems but still there are few challenges to deal with such as recognizing multiwords expressions and linguistically annotating the recognized expressions. Later approaches are discussed in Eisele et al. (2008). Like everything, data-driven techniques have their pros and cons. Though, they help in filling the dictionary gaps by reducing the amount of out-of-vocabulary words, however, as a side-effect they aggravate the problem of lexical selection in already large dictionaries. The increasing problem of selection between the given choices becomes more difficult than solving the problem of dictionary gaps.

Pre-Editing technique lacks in improving MT quality significantly when it comes to applying automatic grammar learning approaches. These approaches suffer due to extraction of large quantity of rule candidates from even small corpora. Due to the noise produce by the grammar extraction techniques, selection of a grammar rule from a huge set of possible candidates often lead to the problems of combinatorics and un-expected side-effects, such as parse failures.

Another approach of using data driven techniques in rule-based systems includes modification of the core engine of the RBMT. This has been tried in several respects such as using probabilistic parser in MT. The focus of current hybrid approaches is more towards translation selection in the transfer phase. Beside existing traditional approaches of transfer selection, more vigorous techniques for lexical selection need to be developed.

- An evident approach is the lexical selection of the most frequent translation of the given expression by disregarding the contextual information.
- Another approach utilizes the contextual information for disambiguating the lexical selection process. The process is based on contextual disambiguation: For a given candidate translation, the context of the given source language is extracted from the corpus during training. Sentences are later translated by matching their context with the context stored during the training. Thurmair (2006) reports the problem of retrieving similar translation results when translating a source sentence with two different contexts. He suggests to expand the context beyond the sentence boundaries. Kim et al. (2002) reports the improvement in accuracy when scope of the context is extended from sentence to paragraph.

Core modifications in RMT can improve the transfer selection process significantly; however they are less successfull in case of robustness and parse failures (Thurmair, 2009).

2.2 Multi-Engine Hybridization (MEH)

Multi-Engine architecture, also referred as Coupling by Thurmair (2009), combines two or more existing systems to produce improved MT output. Combining MT engines together have a long tradition, starting perhaps with Frederking (1994). Multi-engine systems can be roughly divided into two approaches:

The first and simple approach is to select the best output from a number of

systems but leave the individual hypotheses as it is e.g. Hildebrand and Vogel (2008) search for the best n-grams in all output hypotheses available, and then select the best hypothesis from the candidate list. They report an improvement of 2-3 BLEU compared to the best single system, as the resulting text can integrate sentences from different MT system outputs.

The second and more sophisticated approach is to recombine the best parts from multiple hypotheses into a new utterance that can be better than the best of the given candidates. This approach does not work on whole sentences but on smaller segments (phrases, words). It uses confusion networks, and generates an output sentence on the basis of the available MT outputs. One such phrase-level approach is to extract sentence-specific phrase translation tables from system outputs with alignments to source and running a phrasal decoder with this new translation table (Rosti et al., 2007).

3 On going research and work plan

This section presents the main research objectives and some preliminary work already achieved. Moreover, a time scheduled is proposed for the goals to be accomplished.

3.1 Exploring factor-based models for SMT

3.1.1 Experiments for English-to-Czech

In the preliminary work (Bojar et al., 2012), a taxonomy of factored models is explored for English to Czech. The taxonomy is based on the number of translation steps and nature of search space. The variations in decoding configurations are associated with different types of expected problems. Single translation step with single independent search is named as direct setups. This type of configuration is likely to suffer from out-of-vocabulary issues (due to insufficient generalization) on

either side. Direct setups are targeted to explicitly model the target side morphology. The experiments reported some improvements from the surface pharse-based SMT only when the language models of the morphological features are used.

Single-Step setups are the second type of configuration in the taxonomy that model scenarios consist of more than one translation steps within a single search. This configuration have a very high risk of combinatorial explosion of translation options and/or of spurious ambiguity. Single-Step configurations used linguistically motivated decodings e.g. lemma based translation is used and the target surface-form is generated using the morphological factors. Alternate decoding paths are included as fallback options. More complex decoding paths hinted improved results and lemma-based language models showed relative importance.

The third type of configuration used a serial approach of combining multiple searches to reduce the complexity of search space, we called this configuration a Two-Step setup. Serially connected setups can lose relevant candidates between the searches, unless some ambiguous representation like lattices is passed between the steps. In our two-step configurations, in the first step the source gets translated to a simplified Czech and in the second step, the simplified Czech gets fully inflected. The details of various two-step configurations are listed in (Bojar et al., 2012).

3.1.2 Experiments for English-to-Urdu

Factored experiments are highly dependent on linguistic resources to enrich the corpora. Urdu is a resource-poor language, linguistic tools such as morphological analyzers, parsers etc that are required for annotating the data for factored setups are scarce. The first step towards running factored setups for Urdu is the collection of the available resources. Jawaid and Bojar (2012) list available resources for Urdu part-of-speech tagging and morphology and also implement an approach to improve the accuracy and coverage of available taggers by tagger voting. We will use Jawaid and Bojar (2012) tool to obtain the part-of-speech tagged corpus for Urdu.

English/Urdu language pair has very little bilingual data available for training.

So far we have collected three different set of corpora that includes a small set of Penn Treebank sentences translated in Urdu by CRULP[1], Emille corpus published by ELRA, and recently Post et al. published a parallel corpus for six indian languages including English/Urdu pair.

After annotating the existing parallel corpora, the next step is to run the taxonomy for factored models as described in (Bojar et al., 2012) on English-to-Urdu.

3.1.3 Extending the taxonomy with reordering factors

The taxonomy for factored models highlighted in the previous sections based only on translation and generation factors. Morphological factors can also help in learning reordering models. The Moses system allows for integration of multiple factors in reordering models that can be beneficial for learning local reordering for both Czech and Urdu.

3.2 Artificial enriching of source language

When translating from morphologically poor (e.g. English) to morphologically rich (e.g. Czech and Urdu) languages, one of the core issues is to generate the missing information that is not explicitly marked in the surface-form of the source text. For example fixed word order languages express case marking through positional system, which cannot be model using only surface information. Factor based SMT can be utilized to resolve this problem, however the decomposition of the phrase translation into several mapping steps creates additional computational complexity. The other approach to tackle this problem is to artificially enrich the source language with the required linguistic information by creating a pseudo language (Goldwater and McClosky, 2005).

Kamran (2012) adopted this technique for English-to-Urdu SMT. He generated pseudo-english using heuristic rules applied on English dependency parse trees. His

[1]Center for Research in Urdu Language Processing (www.crulp.org)

work also highlights the over generation of case markers due to the use of simplified set of heuristic rules. We plan to extend the set of rules with a deeper understanding of the source and target languages. For example, in the case of English-to-Urdu, we need to properly model the split-ergativity (Durrani and Lahore, 2006) exist in Urdu.

3.3 Automatic extraction of syntax based reordering rules

Use of syntax is a potential solution for long distance reordering. The basic idea is to change the word order of the source sentence to make it more similar to word order of the target sentence in a preprocessing step. This preprocessing step is inspired by previous approaches like Xia and McCord (2004), which split translation into two steps:

$$S \rightarrow S' \rightarrow T$$

Where S is the source language sentence which is reordered in the first step according to the word order of target language, resulting in the reordered sentence S'. In the second step the reordered sentence S' is monotonously translated into the target language sentence T.

Preprocessing by reordering the source side sentences in statistical machine translation has proved to be useful for a number of language pairs e.g. Visweswariah et al. (2011) shows significant improvements for Hindi-to-English, Urdu-to-English and English-to-Hindi; Visweswariah et al. (2010) applied reordering for various languages and showed improvements for English-to-Spanish, French and Hindi.

Jawaid and Zeman (2011) applied this technique on English-to-Urdu using hand-written reordering rules. The hand-written rules for English-to-Urdu reordering significantly improves the translation quality, however, the rules listed in Jawaid and Zeman (2011) do not capture all grammatical structures. Extending the hand-written rules is an intensive task and required a lot of human labour.

In contrast to writing grammatical rules by hand, various attempts have been

made to learn the reordering rules automatically using data driven methods (Visweswariah et al., 2010; Zhang et al., 2007; Lavie et al., 2003) that shows promising results. We will adopt a probabilistic rule-extraction algorithm (Yamada and Knight, 2001) for English-to-Urdu.

3.4 Timelines

Work on this thesis should be completed by September 2015, a tentative timeline for the proposed tasks are as follows:

- **Jan to July 2013**
  - Process the required resources for English-to-Urdu language pair.
  - Complete factor based experiments including the modeling of reordering factors.
- **July to Dec 2013**
  - Work on linguistically motivated rules for generating pseudo-english.
  - Experimentation using pseudo-english for both Czech and Urdu.
- **Jan to July 2014**
  - Acquiring the required knowledge for probabilistic reordering rule-extraction.
  - Implementation and reordering experiments.
- **July to Dec 2014**
  - Applying evaluation techniques and perform error analysis
- **Jan to Sep 2015**
  - Writing of the thesis

References

Jeff A Bilmes and Katrin Kirchhoff. Factored Language Models and Generalized Parallel Backoff. In *Proceedings of HLT-NAACL*, Edmonton, 2003.

Ondřej Bojar and M Prokopová. Czech-English word alignment. *Proc of LREC*, 2006.

Ondřej Bojar, Bushra Jawaid, and Amir Kamran. Probes in a taxonomy of factored phrase-based models. In *WMT '12: Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, June 2012.

Peter F. Brown, Peter V deSouza, Robert L. Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18, 1992.

Michael Collins, Philipp Koehn, and Ivona Kučerová. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, June 2005.

Factored Language Models Tutorial, February 2008. Department of EE, University of Washington.

N Durrani and F N Lahore. System for Grammatical relations in Urdu. *Nepalese linguistics*, 22:91, 2006.

Andreas Eisele, Christian Federmann, Hans Uszkoreit, Herve Saint-Amand, Martin Kay, Micheal Jellinghaus, Sabine Hunsicker, Teresa Herrmann, and Yu Chen. Hybrid machine translation architectures within and beyond the EuroMatrix project. In *Proceedings of the 12th annual conference of the European Association for Machine Translation (EAMT 2008)*, pages 27–34, Hamburg, September 2008.

Robert Frederking. Three heads are better than one. In *Proceeding of the Fourth Conference on Applied Natural Language Processing Stuttgart*, 1994.

Sharon Goldwater and David McClosky. Improving Statistical MT through Morphological Analysis. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

Nizar Habash, Bonnie Dorr, and Christof Monz. Symbolic-to-statistical hybridization: extending generation-heavy machine translation. *Machine Translation*, 23 (1):23–63, November 2009.

A S Hildebrand and Stephan Vogel. Combination of machine translation systems via hypothesis selection from combined n-best lists. *8th AMTA conference*, 2008.

Bushra Jawaid and Ondřej Bojar. Tagger Voting for Urdu. In *In Proceedings of the 24th International Conference on Computational Linguistics, COLING*, Mumbai, 2012.

Bushra Jawaid and Daniel Zeman. Word-Order Issues in English-to-Urdu Statistical Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, (95): 87–106, April 2011.

Amir Kamran. *Hybrid Machine Translation Approaches for Low-Resource Languages*. PhD thesis, Institute of Formal and Applied Linguistics (UFAL), April 2012.

Yu-Seop Kim, Jeong-Ho Chang, and Byoung-Tak Zhang. A comparative evaluation of data-driven models in translation selection of machine translation. In *the 19th international conference*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, University of Edinburgh, 1 edition edition, 2010.

Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondřej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, Alexandra Constantin, Christine Corbett Moran, and Evan Herbst. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. Technical report, September 2007a.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, and Nicola Bertoldi. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, Prague, June 2007b.

Alon Lavie, Stephan Vogel, and Lori Levin. Experiments with a Hindi-to-English Transfer-based MT System under a Miserly Data Scenario. *ACM Transaction on Computational Logic*, 5, September 2003.

Matt Post, Chris Callison-Burch, and M Osborne. Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing. *aclweb.org*.

Chris Quirk, Arul Menezes, and Colin Cherry. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 271–279, 2005.

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

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. Case markers and morphology: addressing the crux of the fluency

problem in English-Hindi SMT. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 800–808, Suntec, August 2009. IIT Bombay, Association for Computational Linguistics.

Antti-Veikko I Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J Dorr. Combining outputs from multiple machine translation systems. In *Proceedings of NAACL HLT*, 2007.

Gregor Thurmair. Using corpus information to improve MT quality. *Third International Workshop on Language Resources …*, 2006.

Gregor Thurmair. Comparing different architectures of hybrid Machine Translation systems. In *MT Summit XII*, 2009.

Kristina Toutanova, Hisami Suzuki, and et al. Applying morphology generation models to machine translation. In *Association of Computational Linguistics*, 2008.

Karthik Visweswariah, Jiri Navratil, Jeffrey Sorensen, Vijil Chenthamarakshan, and Nanda Kambhatla. Syntax Based Reordering with Automatically Derived Rules for Improved Statistical Machine Translation. *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1119–1127, August 2010.

Karthik Visweswariah, Rajakrishnan Rajkumar, Ankur Gandhe, Ananthakrishnan Ramanathan, and Jiri Navratil. A Word Reordering Model for Improved Machine Translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 486–496, Edinburgh, July 2011.

Chao Wang, Michael Collins, and Philipp Koehn. Chinese Syntactic Reordering for Statistical Machine Translation. In *Proceedings of the EMNLP-CoNLL*, 2007.

Fei Xia and Michael McCord. Improving a statistical MT system with automatically learned rewrite patterns. In *20th international conference on Computational Linguistics*, 2004.

Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, 2001.

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

Dongdong Zhang, Mu Li, Chi-Ho Li, and Ming Zhou. Phrase Reordering Model Integrating Syntactic Knowledge for SMT. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 533–540, Prague, June 2007.