ITEC 810 Project Proposal:

Source-context Features for English-to-Czech Machine Translation

Kamil Kos
41903676

Supervisor: Mark Dras

August 21, 2009

Abstract

Statistical machine translation (MT) has become one of the major development streams in MT in recent years. Models taking advantage of source context information have shown that they can improve translation quality. In this project, we investigate the impact of source-context features on the quality of English-to-Czech machine translation. The context we consider are surrounding words and part-of-speech tags, local syntactic structure and other linguistically motivated features that can be extracted from the source language sentence. We implement an extension to the open source MT system Moses, which is used as baseline for our experiments.

Keywords: Machine translation, source context, Moses

1 Project Description

1.1 Background

Language belongs to the most complex systems that mankind has ever created. As a means of communication, it must be precise and concrete enough but vague and inaccurate at the same time to express one's feelings and emotions. It takes several years for children until they learn how to use their own mother tongue correctly. If they want to translate between two different natural languages, e.g. English and Czech, several more years at school learning the other language are required until they are capable to translate a sentence from one language to the other one without a mistake. Machine translation (MT) tries to eliminate the need for spending time learning other languages by providing methods that make it possible to automatically translate from one natural language into another one.

Various approaches to MT were proposed in the past years ranging from simple human-written translation rules to complex linguistically motivated translation systems. As more computational power became available, statistical approaches gained on significance. Their biggest advantage lies in their universal applicability and fast deployment. Within a few days or even hours, it is possible to train a translation system for two arbitrary languages. The only requirement is a large collection of text in both languages that are aligned sentence by sentence. Obtaining such large collections of text, called corpora, is becoming easier since there are vast amounts of text on the internet that can be used for training purposes after careful extraction.

One of the first statistical MT systems was described by [BCP$^+$90] who introduced the word-based translation model. Words in the source sentence can be reordered and than translated using translation probabilities estimated from parallel corpora. The fluency of the produced target text is than checked by the language model, which controls whether the translation in the target language is a well-formed sentence. Language model is trained on large monolingual corpora and it can evaluate which sequences of words are more probable in the target language and which are less probable. As the field of MT evolved, more complicated models were introduced. However, the basic paradigm of translation model and language model are still present in the state-of-the-art statistical MT systems.

1.2 Aims, Significance and Expected Outcomes

One of the successful approaches to automatic translation between natural languages is phrase-based MT which is just another incarnation of the statistical MT

described in the previous section. The major difference to the model introduced by [BCP+90] is that it takes advantage of translating whole phrases instead of single words. Phrases that have a fixed meaning, e.g. "good morning", usually have a fixed translation in the target language. Therefore, they can be translated as one piece. This helps to increase the translation quality because we do not need to translate them word by word. However, the length of the phrases is usually only a few words and the translation of a sentence must be combined from many separate phrases.

A serious issue during the translation process is to decide which translation of the source phrase to choose if there are more options. Since we do not usually know which one is the best, we take all possible translations and create a pool of translation hypotheses. From these hypotheses we select the one which we think is the most probable according to our translation and language model.[1] However, there can be up to hundreds or thousands of possible translations of one phrase in a realistic scenario. Because long sentences consist of many short phrases, considering all combinations of all possible phrase translations is intractable due to the combinatorial explosion. Therefore, we must discard the less probable translations without even looking at them. This represents a problem that has not been sufficiently solved yet. It is possible that we discard a translation because we think it is not probably the correct translation but it can actually be the preferred translation in a given context.

To tackle this problem, source-context features have been introduced [CW07, GS08]. They try to provide additional information about the context in which the phrase translations are usually used. This helps to better distinguish among the translation options since each of them can be used in a slightly different situation.

Because the source sentence is fully observable during the translation, i.e. we know which sentence we are translating, a full range of features can be implemented without any need of changing the decoding[2] algorithm of phrase-based MT. Therefore, it is easy to implement an extension of the existing MT systems just by adding several additional features and the translation time should not be increased significantly.

In [CW07] and [GS08], following source-context features were proposed:

- words that co-occur in the sentence,
- part of speech tags surrounding the phrase,
- basic dependency features,
- position of the phrase in the sentence.

Both papers reported improvements of the translation quality for the Chinese-to-English translation. Slight improvements were also reported for the English-German language pair in both directions but the improvement was not as high as for the Chinese-English pair.

In this project, we would like to investigate the influence of the source-context features on the translation pair English-Czech with Czech being the target language. Because the improvement of MT quality was not the same for German and Chinese we would like to measure the extent to which source-context features can improve English-to-Czech translation and identify the most suitable ones.

[1] State-of-the-art systems contain several additional models to improve their performance, but we will consider only the translation and language model for the sake of simplicity.

[2] The term decoding denotes the translation procedure.

Czech belongs to the Slavic language family together with Russian, Polish and other East European languages. The results we obtain in this project will probably hold also for other languages from this family. If future research shows that source-context features improve the translation quality of an arbitrary language, they can be incorporated into translation systems as another development step towards their better performance.

2 Methodology and Plan

2.1 Approach

We decided to use the phrase-based statistical MT system Moses [Koe09] as baseline system for our experiments. It is a translation system that is widely used for research purposes and its performance in terms of translation quality is competitive with state-of-the-art commercial MT systems.

Moses implements the log-linear model in which a source-language sentence f is translated into the target-language sentence $\hat{e}$ that maximizes a linear combination of features h_n and their corresponding weights λ_n. This can be expressed by the following equation:

$$\langle \hat{e}, \hat{a} \rangle = argmax_{\langle e,a \rangle} \sum_{n=1}^{N} \lambda_n h_n(e, a, f)$$

where a is the segmentation of e and f into phrases and h_n are feature functions that return logarithm of the feature probability. Because the model uses a linear combination of features that compute probability logarithms, it is called log-linear.

The model is easily extensible by additional features that can be added without significant modifications to the framework. The weights λ_n are automatically optimized to an objective function that minimizes the translation errors. Therefore, it is sufficient to provide new features for which the weights are automatically learned. Our goal is to implement the source-context feature functions mentioned in Section 1.2 since the log-linear model will adapt itself to them after the weights-optimization phase.

2.2 Task Plan

We have divided our plan into five tasks. The estimated time requirements for each of the tasks are depicted in Figure 1.

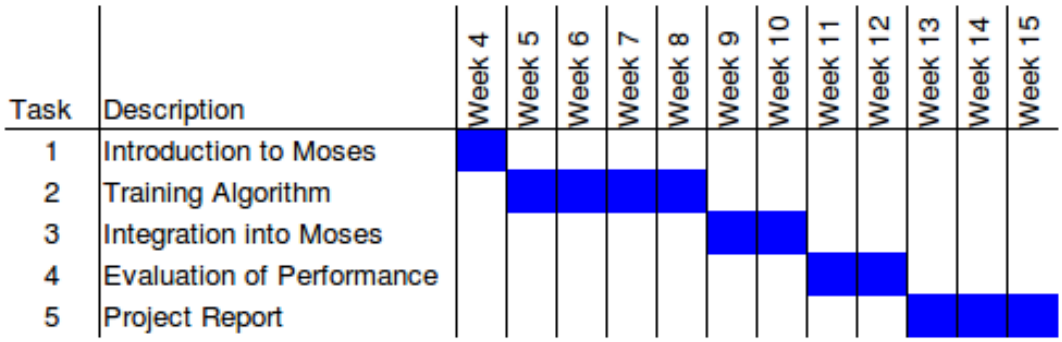

Figure 1: Project schedule

2.2.1 Task 1 - Introduction to Moses

In this preparatory phase, we will thoroughly read the documentation of the Moses MT system in order to have sufficient knowledge about the internal implementation of Moses. This will help us in the later phases to incorporate effectively our extension into the translation system. We estimate that we will need one week for this task.

2.2.2 Task 2 - Training Algorithm

This is the core part of our project, in which we want to design and implement the training algorithm for source-context features. They will be automatically extracted from parallel corpora and stored in a dedicated data structure that will allow efficient manipulation with them. We will implement training algorithm for features mentioned in Section 1.2. The duration of this phase will be four weeks.

2.2.3 Task 3 - Integration into Moses

After we finish the training algorithm for source-context features, we will integrate them into the decoding algorithm of Moses, so that they are taken into account during the translation process. At the end of this phase, which will require two weeks of work, we will be able to run Moses with our extension translating sentences from English to Czech.

2.2.4 Task 4 - Evaluation of Performance

In order to see the impact of our extension on the translation quality, we will evaluate it against baseline Moses without our source-context extension using the most common MT metrics, e.g. BLEU [PRWjZ02]. We will evaluate the change in performance on official testsets that are used by the MT community. We plan to run the evaluation on WMT08 and WMT09 data (Workshop on Statistical Machine Translation). The estimated time for this task are two weeks.

2.2.5 Task 5 - Project Report

We will describe our work and results in an extensive report. We plan to comment on the contribution of the individual source-context features to the translation quality for the language pair English-Czech, and we will compare our results to findings obtained for other languages published in the literature. The duration of this phase will be three weeks.

References

[BCP+90] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, *A statistical approach to machine translation*, Computational Linguistics **16** (1990), no. 2, 79–85.

[CW07] Marine Carpuat and Dekai Wu, *Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation*, Proceedings of Machine

Translation Summit XI (Copenhagen, Denmark), 2007, pp. 73–80 (english).

[GS08] Kevin Gimpel and Noah A. Smith, *Rich source-side context for statistical machine translation*, Proceedings of the Third Workshop on Statistical Machine Translation (Columbus, Ohio), Association for Computational Linguistics, June 2008, pp. 9–17.

[Koe09] Philipp Koehn, *Moses - a Beam-Search Decoder for Factored Phrase-Based Statistical Machine Translation Models*, University of Edinburgh, August 2009.

[PRWjZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu, *BLEU: a Method for Automatic Evaluation of Machine Translation*, 2002, pp. 311–318.