ITEC 810 Project Outline:

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

Kamil Kos
41903676

Supervisor: Mark Dras

October 16, 2009

Contents

1 Machine Translation **2**

2 Past Work **4**

3 Proposed Source-context Features **5**

4 Methodology **6**

5 Results and Discussion **6**

Abstract

Machine translation (MT) uses computers for translation between natural languages, e.g. English and Czech. In recent years, statistical MT has become one of the major development streams in MT because it can be adopted to any arbitrary language pair fairly fast without special linguistic knowledge of the languages. It uses large mono- and bilingual texts from which translation rules are automatically derived by advanced statistical methods. Models taking advantage of source context information have recently 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. Source-context features use additional information from the sentence to be translated to better disambiguate between different translation options. The context we consider are surrounding part-of-speech tags, local syntactic structure and position of a phrase in the sentence. We implement an extension to the open source phrase-base MT system Moses, which is used as baseline for our experiments.

Keywords: Machine translation, source context, Moses

1 Machine Translation

Before we start with the description of our project, we shortly introduce the field of machine translation (MT) and mention several important milestones that influenced the development of the field. We focus especially on statistical MT, which represents one of the fastest developing streams in MT and which has been receiving a lot of the research attention in recent years. Thanks to this fact, statistical MT has been able to report steady improvement in translation quality in the last decade and is capable of translating even long sentences relatively accurately. However, flawless automatic translation between natural languages is still an issue for future research.

Statistical MT was pioneered by IBM researchers [Brown et al., 1990], who based their work on the "noisy-channel" model. The idea of using the noisy-channel model dates back to the 1950's when it was introduced in the information theory to retrieve the original message from data sent over an unreliable and noisy transmission channel. The noisy-channel model views the source text as encoded target language text. In the following, we will use the terms source and target sentence in the MT fashion:

source denotes the language from which we translate,

target denotes the language to which we translate.

Note that source and target side are used in the opposite way in information theory while speaking about the noisy-channel model.

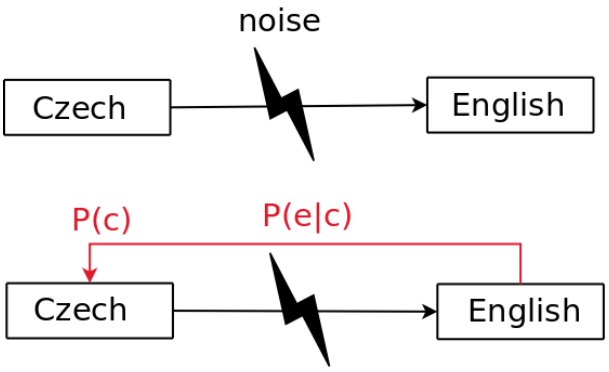

Figure 1: The noisy-channel model combines the language model $P(c)$ and the translation model $P(e|c)$ to find the best Czech translation of an English sentence.

In the following section, we discuss the noisy-channel model in detail, which represents the basic model of statistical MT. It consists of the translational model, expressed by the conditional probability $P(e|c)$, and the language model, expressed by the probability $P(c)$. These two parts are the basic constituents of the noisy-channel model, which is depicted in Figure 1.

Training of a statistical MT system is one of the most computationally extensive parts in the construction of a MT system. It requires the estimation of all translation and language model probabilities as accurately as possible from large amounts of data. The major problem is that training data represent only a subset of all possible sentences in a language. Even if we collect as much training data as possible, we will still miss a lot of grammatically correct sentences that will not appear in our training data. Therefore, it is not possible to get a sufficient amount of information about the actual real-world probabilities and we can only compute rough estimates. However, this is often enough to construct a reasonably well performing statistical MT system.

While training of a MT system requires a lot of computational power to estimate all necessary parameters of the underlying model and can take several hours up to several days of intensive parameter optimisation, the actual translation of a sentence is more time critical. We do not want to wait for translation of one sentence for hours or even days. Therefore, approximation methods have to be used in order to compute the resulting translation fast enough. The process of translation is called decoding in statistical MT. It relates to the noisy-channel model in which we try to extract the original sentence from a corrupted text, i.e. to "decode" the sentence. In this section, we describe the decoding procedure and how translational and language model are combined to get the best possible translation of a sentence.

In this section, we shortly cover linguistic foundations that are required to understand source-context features. We describe what part-of-speech (POS) tags are and present their variations used for English and Czech. We also

present the two basic notations that are used to capture the syntax structure of a sentence: phrase structure and dependency structure notation. Finally, we introduce the combinatorial categorial grammar (CCG) and its advantages for capturing sentence structure.

2 Past Work

There are two ways to improve the performance of a statistical MT system. First, it is possible to produce even larger collections of mono- and bilingual texts to get better coverage of the language and estimate the translation and language model more accurately. This approach is easy to do but it requires a lot of time and human effort to select suitable sources of bilingual training data that can be aligned on the sentence level. Usually, it is easy to find parallel texts for a specific domain, e.g. legal documents produced in states or organisations that have multiple official languages. However, for some other domains like newspaper articles it is difficult to find a suitable source of data because they are usually produced only in one language. Another approach to improve the quality of statistical MT systems is to enhance the model by additional features that help to overcome the data sparsity problem. Source-context features belong to this category. In this section, we describe the past research on source-context features.

First, we introduce the work described in [Carpuat and Wu, 2007], in which the authors incorporated a word sense disambiguation (WSD) module into the decoding procedure in order to achieve better phrasal lexical choice. They used a phrase-based decoder Pharaoh to compute the baseline scores and augmented it with the WSD module. They reported an improvement of the translation quality on a Chinese-to-English translation task using context-dependent phrasal translation lexicon to take advantage of the additional source-context information.

A similar approach was reported in [Gimpel and Smith, 2008]. They used a wider set of source-context features than [Carpuat and Wu, 2007], including lexical, syntactic and positional features. However, the results were not so convincing as in the work mentioned in the previous paragraph. They report improvement for Chinese-to-English translation and slight improvement for English-to-German translation. However, no statistically significant results could be measured for German-to-English translation tasks. Since we are investigating the influence of source-context features on the English-to-Czech direction and Czech has rich morphology as German does, we expect that the source-context features can have similar contribution to the translation quality as for the English-to-German translation.

Combinatorial categorial grammar (CCG) is a formalism that enables to tag individual words with their syntactical relations. This property was used in [Birch and Osborne, 2007] as additional source-context information. The

results of experiments conducted on Dutch-to-English translation indicate that this feature contributed the most to correct reordering of words in the sentence. This feature would not be as useful for English-to-Czech translation since Czech has a relatively free word order. Therefore, we do not include it in our feature set.

3 Proposed Source-context Features

In this section, we present selected features that we decided to implement as part of our project. First, we provide a short description of each of the features and give examples at situations where they can contribute to better translation. Most of the features are inspired by past work on source-context features mentioned in the previous section. However, we would like to see if the quality improvement is language dependent and the for each language is better a different set of source-context features.

As reported in [Gimpel and Smith, 2008], the most useful source-context feature was the surrounding part-of-speech (POS) tags of a phrase. This feature uses shallow syntax to disambiguate individual translation options. It requires that the sentence to be translated is analysed by a tagger and all words need to be assigned a POS tag. The impact of this feature is a little bit similar to the function of the language model - they both check that the translated sentence contains only words that have been seen in the training data with high probability. However, the language model works only on the level of words while this proposed feature can work with longer phrases, too. Furthermore, the language model works on the target side of the language pair whereas we consider surrounding POS tags on the source side. Therefore, this source-context feature provides additional information to the decoder and is not just a variation of the language model.

The next proposed source-context feature are basic dependency features. They require that the source sentence in parsed by a syntactical parser that computes the sentence parse tree. The parse tree captures basic syntactical relations between words in the sentence. These relations can be used to distinguish between multiple translation options. In this section, we describe which basic dependency relations we use as source-context features.

The last proposed feature is the position of the phrase in a sentence. The impact of this feature will most probably not be very high but it can be useful in languages with fixed word order like English. Words or phrases with a special function have a fixed position in the sentence in languages with fixed word order. Therefore, their position in the sentence can have influence on the way how they are translated.

4 Methodology

We decided to use the statistical MT system Moses [Koehn, 2009] as our baseline system. It is an open-source phrase-based decoder that implements the beam-search algorithm to find the best translation of a source sentence. It is widely used by the research community for testing new features for MT. In this section, we describe the decoding procedure that was implemented in Moses more in detail. Moreover, we shortly discuss additional features to the basic noisy-channel model that are implemented in Moses to improve its performance.

We concentrate especially on the translation model that is implemented in Moses. The model is called log-linear because it uses a linear combination of features that compute probability logarithms. It is easily extensible by additional features that can be added without significant modifications to the framework. The weights λ_n of the features 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 3 since the log-linear model adapts itself to them after the weights-optimization phase.

In the following section, we describe the implementation details of our proposed Moses extensions. We use a special data structure called a suffix array to compute document statistics efficiently. We are especially interested in frequency of individual phrases that are extracted from the bilingual training corpus. Since the training corpus can be very large, it would be intractable to store all possible phrases with their probabilities since the context we consider can consume a lot of space. Suffix arrays require only $O(n)$ space for their own internal representation, where n is the size of the training corpus, but they make it possible to compute frequency counts of an arbitrary phrase only in $O(log(n))$ time. This is their biggest advantage because the phrase probabilities do not need to be precomputed but can be extracted on-demand during the decoding step.

5 Results and Discussion

Evaluation of results in an important part of each experiment. There are standard methods how to evaluate MT quality. The most widely used MT metrics are automatic metric that compare the MT output with human reference translation(s) of the same sentence. Such metrics are easy and fast to compute. Some people argue that these metrics do not correlate with human judgements sufficiently enough and provide biased results. However, the speed and cost advantage to human manual evaluation is significant and it beats all other drawbacks. We use the n-gram metrics Bleu

[Papineni et al., 2002] and Nist [Doddington, 2002], which are viewed as the standard metrics in MT, to evaluate the performance of the translation system. Moreover, we also compute the GTM metric [Turian et al., 2003], which combines recall and precision of individual words and rewards sequences of words found both in the candidate and reference hypothesis.

In the following section, we provide results for individual source-context features. We compare their performance to the baseline system and identify features that improved the MT quality. We evaluate the features on two official test sets that were used at the Workshop for Statistical Machine Translation (WMT) in years 2008 and 2009.

Finally, we identify the best performing features or their combinations and try to explain why they performed better than other features. We also compare our results to data published in the recent years for other languages. We discuss why the contribution of features is similar or different for different languages. To conclude, we propose new potential directions for source-context features that showed promising results and could lead to better MT quality.

References

[Birch and Osborne, 2007] Birch, A. and M. Osborne: 2007, 'CCG Supertags in Factored Statistical Machine Translation'. In: *In ACL Workshop on Statistical Machine Translation*. pp. 9–16.

[Brown et al., 1990] Brown, P. F., J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin: 1990, 'A Statistical Approach to Machine Translation'. *Computational Linguistics* **16**(2), 79–85.

[Carpuat and Wu, 2007] Carpuat, M. and D. Wu: 2007, 'Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation'. In: *Proceedings of Machine Translation Summit XI*. Copenhagen, Denmark, pp. 73–80.

[Doddington, 2002] Doddington, G.: 2002, 'Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics'. In: *Proceedings of the Second International Conference on Human Language Technology Research*. San Francisco, CA, USA, pp. 138–145.

[Gimpel and Smith, 2008] Gimpel, K. and N. A. Smith: 2008, 'Rich Source-Side Context for Statistical Machine Translation'. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio, pp. 9–17.

[Koehn, 2009] Koehn, P.: 2009, 'Moses - a Beam-Search Decoder for Factored Phrase-Based Statistical Machine Translation Models'. University of Edinburgh.

[Papineni et al., 2002] Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu: 2002, 'BLEU: a Method for Automatic Evaluation of Machine Translation'. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. pp. 311–318.

[Turian et al., 2003] Turian, J. P., L. Shen, and I. D. Melamed: 2003, 'Evaluation of Machine Translation and its Evaluation'. In: *Machine Translation Summit IX*. pp. 386–393.