A PROTOTYPE ENGLISH-TURKISH STATISTICAL MACHINE TRANSLATION SYSTEM

by

ILKNUR DURGAR EL-KAHLOUT

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
DOCTOR OF PHILOSOPHY

Sabancı University
June 2009

A PROTOTYPE ENGLISH-TURKISH STATISTICAL MACHINE TRANSLATION SYSTEM

APPROVED BY

Prof. Dr. Kemal Oflazer (Thesis Supervisor) .....

Assoc. Prof. Dr. Berrin Yanıkoğlu .....

Assist. Prof. Dr. Hakan Erdoğan .....

Assist. Prof. Dr. Hüsnü Yenigün .....

Assist. Prof. Dr. Deniz Yüret .....

DATE OF APPROVAL: .....

to my little Ahmed

Acknowledgments

I would like to express my gratitude to my supervisor Kemal Oflazer for his guidance, suggestions and especially his patience throughout the development of thesis. I would like to thank all my jury members Berrin Yanıkoğlu, Hakan Erdoğan, Hüsnü Yenigün and Deniz Yüret for reading and commenting on this thesis.

I would like to thank to my all labmates, Özlem, Alisher, Reyyan, Süveyda and Burak. I am grateful to my family for their support and help thoughout my whole life. And I owe a great dept of thanks to my husband Yasser for his helps and encouragements.

This work was supported by TÜBİTAK – The Turkish National Science and Technology Foundation under project grant 105E020. Şeyma Mutlu implemented the word-repair code. Cüneyd A. Tantuğ implemented the BLEU+ tool.

A PROTOTYPE ENGLISH-TURKISH STATISTICAL MACHINE TRANSLATION SYSTEM

Abstract

Translating one natural language (text or speech) to another natural language automatically is known as machine translation. Machine translation is one of the major, oldest and the most active areas in natural language processing. The last decade and a half have seen the rise of the use of statistical approaches to the problem of machine translation. Statistical approaches learn translation parameters automatically from alignment text instead of relying on writing rules which is labor intensive.

Although there has been quite extensive work in this area for some language pairs, there has not been research for the Turkish - English language pair. In this thesis, we present the results of our investigation and development of a state-of-the-art statistical machine translation prototype from English to Turkish. Developing an English to Turkish statistical machine translation prototype is an interesting problem from a number of perspectives. The most important challenge is that English and Turkish are typologically rather distant languages. While English has very limited morphology and rather fixed Subject-Verb-Object constituent order, Turkish is an agglutinative language with very flexible (but Subject-Object-Verb dominant) constituent order and a very rich and productive derivational and inflectional morphology with word structures that can correspond to complete phrases of several words in English when translated.

Our research is focused on making scientific contributions to the state-of-the-art by taking into account certain morphological properties of Turkish (and possibly similar languages) that have not been addressed sufficiently in previous research for other languages. In this thesis; we investigate how different morpheme-level representations of morphology on both the English and the Turkish sides impact statistical translation results. We experiment with local word ordering on the English side to bring the word order of specific English prepositional phrases and

auxiliary verb complexes, in line with the corresponding case marked noun forms and complex verb forms, on the Turkish side to help with word alignment. We augment the training data with sentences just with content words (noun, verb, adjective, adverb) obtained from the original training data and with highly-reliable phrase-pairs obtained iteratively from an earlier phrase alignment to alleviate the dearth of the parallel data available. We use word-based language model in the re-ranking of the `n`-best lists in addition to the morpheme-based language model used for decoding, so that we can incorporate both the local morphotactic constraints and local word ordering constraints. Lastly, we present a procedure for repairing the decoder output by correcting words with incorrect morphological structure and out-of-vocabulary with respect to the training data and language model to further improve the translations. We also include fine-grained evaluation results and some oracle scores with the BLEU+ tool which is an extension of the evaluation metric BLEU.

After all research and development, we improve from 19.77 BLEU points for our word-based baseline model to 27.60 BLEU points for an improvement of 7.83 points or about 40% relative improvement.

Özet

Bir dilin (yazı ya da konuşma) diğer bir dile bilgisayar ile otomatik olarak çevrilmesi bilgisayarlı çeviri olarak bilinmektedir. Bilgisayarlı çeviri doğal dil işlemenin çok eskiden bu yana ilgilendiği en önemli ve aktif konulardan biridir. Son bir kaç on yılda bilgisayarlı çeviri probleminde istatistiksel yaklaşımların kullanımında artış gözlenmiştir. İstatistiksel yaklaşımlar sembolik yaklaşımlardan daha basit olmalarına rağmen yaklaşık sonuçları hiçbir dilbilimsel bilgiye ihtiyaç duymadan üretebilir. İstatistiksel yaklaşımda amaç, sistem parametrelerinin çok fazla zaman ve insan gücüne ihtiyaç duyan, elle yazılan kurallar yerine otomatik olarak öğrenilmesidir.

İstatistiksel bilgisayarlı çeviri bir çok farklı dil çiftleri için uygulansa da, bu alanda Türkçe - İngilizce dil çifti için bir araştırma ve geliştirme çalışması bulunmamaktadır. Bu tezde, İngilizce'den Türkçe'ye en gelişkin istatistiksel bilgisayarlı çeviri prototipinin araştırma ve geliştirilmesin sonuçları sunulmaktadır. İngilizce'den Türkçe'ye istatistiksel bilgisayarlı çeviri prototipi geliştirilmesi bir çok açıdan dikkate değer bir problemdir. En zorlayıcı kısmı, İngilizce ve Türkçe'nin tipolojik olarak görece uzak diller olmasıdır. İngilizce çok limitli bir morfolojiye ve görece sabit bir Özne-Fiil-Nesne öğe sıralamasına sahipken, Türkçe İngilizce'ye çevrildiğinde bir çok sözcüklü öbeğe karşılık gelen sözcük yapılarına sahip, çok zengin ve üretken türetim ve çekimli bir morfolojisi olan çok esnek (Özne-Nesne-Fiil egemen olmakla beraber) öğe sıralamalı eklemeli bir dildir.

Araştırmamız başka diller için yapılan önceki araştırmalarda yeteri kadar çalışılmamış, Türkçe'nin morfolojik özelliklerini dikkate alarak son bilgisayarlı çeviri teknolojisine bilimsel katkılar yapmaya odaklanmıştır. Bu tezde; Hem İngilizce hem de Türkçe tarafında morfolojinin morfem seviyesindeki farklı gösterimlerinin istatistiksel çeviri sonuçları üzerinde nasıl etki yaptığını inceledik. Sözcük eşleşmelerine

yardımcı olmak için, Türkçedeki isim formları ve karmaşık fiil formlarını ile aynı sözcük sıralamasında olması için İngilizce tamlama ve yardımcı fill komplekslerinde lokal sözcük sıralaması deneyleri yaptık. Var olan paralel metinlerin azlığını hafifletmek için, eğitim verisini hem orjinal veriden elde edilen içerik sözcükler (isim, fiil, sıfat, zarf) ile hem de tekrarlı olarak bir önceki sözcük öbeği tabanlı sözcük eşleşmelerinden elde edilen yüksek güvenilirlikli sözcük öbeği çiftleri ile arttırdık. Çözümleme için kullanılan morfem bazlı dil modeline ek olarak n- en iyi listelerini yeniden skorlaması için sözcük bazlı dil modelini kullandık, böylece hem lokal morfotaktik kısıtlamaları hem de lokal sözük sıralaması kısıtlamaları üzerine çalıştık. Son olarak çevirileri, iyileştirmek amacıyla eğitim verisi ve dil modeline göre sözcük dağarcığının dışında olan ve morfolojik yapısı hatalı olan çıktının sözcüklerini onarmak için bir prosedür sunduk. Ayrıca BLEU değerlendirme metriğinin bir uzantısı olan BLEU+ aracı ile elde edilen detaylı değerlendirme sonuçlarını ve elde edilebilecek en yüksek skorlardan bazılarını ekledik.

Tüm araştırma ve geliştirme sonucunda 19.77 BLEU skoru olan sözcük bazlı temel modelimizi 7.83 BLEU skoru ya da %40'lık artışla 27.60 BLEU skoruna geliştirdik.

Contents

Chapter 1

INTRODUCTION

Translating one natural language (source language) to another natural language (target language) automatically is known as machine translation (MT). Machine translation is one of the major, oldest and still the hottest topics in natural language processing research. Translation comprises analysis of the source language sentence, an optional transfer step and generation of the target language sentence. Analysis attempts to extract the structure and the meaning of the source sentence while transfer and generation create an equivalent target language sentence from output of analysis.

Machine translation problem was introduced by Warren Weaver [1] in 1949. He describes the translation process as a cryptography problem: A text written in Russian can be seen as a text written in English but with some *different* symbols. The task is to learn the encryption rules to obtain from the observed text.

Direct dictionary lookup approaches are not sufficient for finding these rules when we talk about translating natural languages. Languages are very complex and the same meaning can be expressed in many different ways. There is rarely a word-to-word correspondence between any two languages so translation can never be seen as a straightforward procedure.

For a successful/accurate translation, a translator should "know" both languages; possess an understanding of their grammars, syntax, semantics, writing conventions, idioms, etc. and moreover take into account the context of source language. This task is easier for a human translator but extremely hard for a computer (at least for now).

First attempts for an English to Turkish machine translation system prototype started in the 1980's [2]. In the 1990's two different English to Turkish machine translation systems [3, 4] were developed as a part of the TU-LANGUAGE project supported by NATO Science for Stability Program. Both systems were rule-based and implemented by manually writing a large number of transfer and generation rules. These systems took advantage of very specific domains (broadcast news captions and IBM computer manuals) with limited context and limited lexical ambiguity.

1.1 Motivation

The latest and most popular machine translation paradigm in the last twenty years is statistical machine translation, which relies on developing statistical models of the translation process from large amounts of parallel data. The main idea is to find the most probable translation for a given sentence by using this statistical model of translation. Thus the intensive human labor for writing transfer and generation rules of previous approaches is replaced by a machine learning process. We review the statistical machine translation paradigm, its methods and challenges in Chapters 2 and 3.

Although there has been quite extensive work in statistical machine translation for some specific language pairs, there has not been any research and development efforts for Turkish - English language pair. The challenges, such as limited data, rich morphology of Turkish, word order, tense differences of English and Turkish,

have been the main motivation of English to Turkish statistical machine translation research. This thesis presents an English to Turkish statistical machine translation system prototype that is the first attempt for this language pair. Our aim in this line of work is to develop a comprehensive model of statistical machine translation from English to Turkish.

Initial explorations into developing a statistical machine translation system from English to Turkish point out that using standard models and techniques to determine the correct target translation is probably not a good idea. The main aspect that would have to be seriously considered first is the Turkish productive inflectional and derivational morphology. A word-by-word alignment between an English-Turkish sentence pair has some Turkish words aligned to whole phrases in the English side, as embedded Turkish morphemes are translated to surface as English words. Thus for an accurate word alignment, we need to consider sublexical structures i.e., parts of words. The details of the model have to at least take into consideration a probabilistic model of the morpheme sequencing in addition to models of higher level word order. This will certainly require certain non-trivial amendments to the translation models developed so far for various other language pairs.

There has been some recent work on translating to and from Finnish (agglutinative language, similar morphological structure with Turkish) in the Europarl corpus [5]. Reported *from* and *to* translation scores for Finnish are the lowest on average over 11 european languages, even with the large number of sentences available. These may hint at the fact that standard alignment models may be poorly equipped to deal with translation from a poor morphology language like English to a complex morphology language like Finnish or Turkish.

1.2 Contributions of the Thesis

This thesis presents the results of an English-to-Turkish phrase-based statistical machine translation study. This language pair is interesting for statistical machine translation for a number of reasons. Most challenging one is that English and Turkish are typologically rather distant languages. English has very limited morphology and rather fixed Subject-Verb-Object constituent order, while the target language, Turkish, is an agglutinative language with very flexible (but Subject-Object-Verb dominant) constituent order and a very rich and productive derivational and inflectional morphology with infinite vocabulary.

The major results of our work can be summarized as follows:

- We experiment with different morpheme-level representations for English - Turkish parallel texts with different derivational morpheme groupings in the Turkish texts.
- We experiment with local word ordering on the English side to bring the word order of specific English prepositional phrases and auxiliary verb complexes, in line with the corresponding case marked noun forms and complex verb forms on the Turkish side to help with alignment.
- We also augment the training data with sentences composed of just content words that are obtained from the original training data to bias content word alignment, and with highly-reliable phrase-pairs from an earlier corpus-alignment.
- We use word-based language model in the re-ranking to generate the n-best lists besides the morpheme-based language model used for decoding.
- Lastly, we present a scheme for *repairing* the decoder output by *correcting* words with incorrect morphological structure and out-of-vocabulary with re-

spect to the training data and language model to further improve the translations.

- We also presented our discussions about the experiments with BLEU+ [6] tool, based on BLEU metric, with some extensions for fine-grained evaluation of morphologically complex languages like Turkish.

We improve from 19.77 BLEU [7] points for our word-based baseline model to 27.60 BLEU points, about 40% increase.

1.3 Outline

The outline of the thesis is as follows:

Chapter 2 starts with a brief history of machine translation. We introduce the basic idea behind statistical machine translation (SMT). We then describe various approaches to SMT such as word-based, phrase-based and factor-based models. We also describe the decoding process and how results are evaluated.

Chapter 3 presents the motivation and challenges of English-to-Turkish statistical machine translation. We analyze data issues, alignment problems and the morphological, grammatical and syntactic contrasts of the languages. We explain why we cannot utilize the state-of-the-art models in English-to-Turkish statistical machine translation and describe a detailed analysis our proposal about morphology integration. Lastly, we explain the preprocessing applied to data and conclude with corpus statistics.

Chapter 4 defines several experiments for a more accurate English-to-Turkish statistical machine translation. These experiments include different morphemic representation schemes with Turkish specific segmentations, content word augmentation to effectively use the training data, English derivational morphology segmentation and local reordering of English phrases to obtain a more monotone alignments. We

conclude the chapter with experimental setup, detailed analysis of experimental results and some examples from the translation of the test data.

Chapter 5 explains our post-processing steps on the decoder output by phrase table augmentation and word repair on the malformed and out-of-vocabulary words. We describe the experimental setup and present our results with a summary of all findings. We also include a fine-grained evaluation results and some oracle scores with the BLEU+ tool which is an extension of the evaluation metric BLEU.

Contributions and future work follow in Chapter 6.

Chapter 2

STATISTICAL MACHINE TRANSLATION

The first and main goal of machine translation is to develop fully automatic high quality machine translation systems. However, research in the past sixty years showed that this goal is not easy to achieve except in very restricted domains. MT systems usually generate outputs that just give the rough meaning and should be post-edited by human translators.

Machine translation systems are differentiated along two dimensions: These are (i) the analysis and generation depth and (ii) the level at which transfer is done. Figure 2.1 shows the Vauquois triangle defining levels of translation. In *direct translation*, the components of source text (words, phrases, etc.) are translated directly without any deep analysis and additional representation. Only very low level of analysis that is very crucial is allowed such as morphological analysis and disambiguation, very local word order changes etc.

In *transfer-based approaches*, analysis and generation are performed before and after transfer. The intermediate representation generated by analyzing source language is transformed to an abstract target representation by using the so-called transfer rules. The target text is generated by using the target specific generation

rules.

The *interlingual approach* is very similar to transfer-based approach except that this level does not have a transfer phase. The interlingua approach uses just one abstract representation scheme which is language independent. So only analysis and generation are sufficient. However, a proper and complete representation which is language independent is very hard to attain.

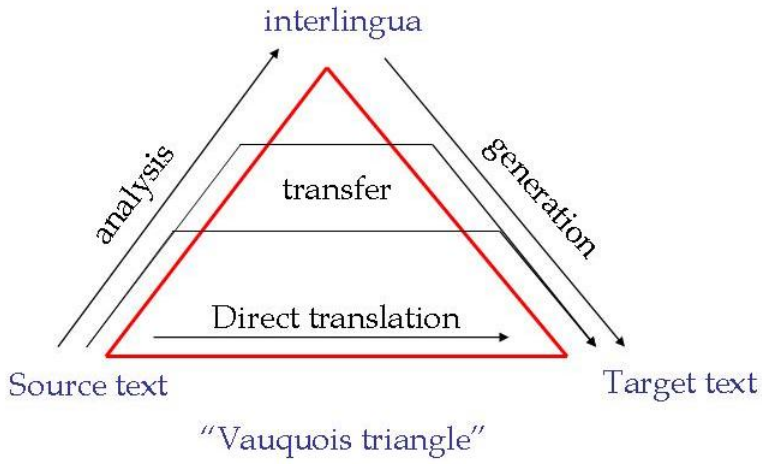


Figure 2.1: Vauquois Triangle

Languages on which machine translation efforts are concentrated show variations during time and are shaped mostly by bussiness and political needs. The first popular language pair was Russian-English in the post World War II period. French-English has also been one of the most studied language because of the bicultural structure of Canadian parliament. European languages gained importance in machine translation research as the translation needs of European Uninon increased due to operational reasons. Arabic-English and Chinese-English are the most popular language pairs due to mostly political and bussiness needs with less prominent efforts on other language pairs.

2.1 A Brief History of Machine Translation

The history of machine translation starts in the 1950s (just after the World War II) with the Georgetown Experiment [8]. In this work, IBM researchers succeeded to translate over sixty Russian sentences into English full automatically by using 250 words and 6 rules. This experiment was a great success and got many researchers interested in MT. Dominating machine translation paradigm in this period was the rule-based approach.

Unfortunately, for many years following the Georgetown experiment, no serious success or improvement was observed which lead to the publication of the ALPAC report [9] in 1966. This report caused a big decline in machine translation research especially in US, claiming that the progress was very far away to fulfill the expectations. However work in Canada and Europe continued. One of the first successful applications was Meteo system that translates weather forecasts from English to French and vice versa till the 1990's. At the same time the first roots of most famous and successful rule-based SYSTRAN started to develop. SYSTRAN is a multilingual machine translation system using direct translation approach and now translates between more than 20 languages. It was used in search engines such as Google and still being used in AltaVista's Babel Fish and global agencies such as NATO, European Union.[1]

Although the rule-based approaches work fine for limited/specific domains, it has many deficiencies. For a wide domain, they need extensive number of manually hand-written rules and lexicons, which is very time consuming, to build. These rules depend on the source and target language, should be written for each language pair and cannot be easily generalized to any other language pair. Moreover, for large domains, the definition of intermediate representation or interlingua is very hard to describe. As a result, rule-based approach is considered improper for general purpose machine translation.

[1]Google moved to its own statistical MT system in 2007

2.2 The Statistical Approach

Following the lack of success of the earlier symbolic or rule-based approaches in developing wide-coverage machine translation systems, [8], the availability of large amounts of parallel electronic texts and increase in the computational power have motivated researchers to shift from rule-based to corpus-based paradigms. The first approach that uses parallel texts as knowledge base is the example-based approach proposed in mid-1980's. The example-based approach treats the corpora as the set of translation examples. Word and phrase translations are selected from analogous examples at run-time. Translation procedure contains decomposition of source texts into segments, searching for matching pre-analyzed phrases of the source language corpus, selecting equivalent target phrases and lastly combining these phrases together to build the target text steps [10]. As generation is done with phrases from actual translations, the target text is more accurate and can deal with language specific idioms and proverbs. The main disadvantage of example-based MT is the need for large parallel corpora for high quality translations.

The major paradigm in the last twenty years in machine translation has been *statistical machine translation* (SMT) which started with the seminal work at IBM [11, 12]. It is still a very active research area. The effectiveness of this paradigm has made a big impact on the MT community as intensive human labour for writing transfer and generation rules is replaced with the statistical methods which are automatic, fast and easier to implement. Moreover statistical approaches usually perform better than the earlier approaches with much less human effort.

The first statistical machine translation approach was IBM's purely statistical word-based model [11, 12]. Experiments on SYSTRAN and IBM's machine translation system (CANDIDE) showed that statistical methods surpass rule-based approaches [13] and they have a great advantage in adapting systems for new domains easily. In the early 2000's, the state-of-the-art translation unit became word phrases instead of individual words [14–17] and very recently, factors have been used as

translation units [18].

In general, any standard statistical machine translation system comprises three components: A training data composed of well-formed and grammatical sentences, a learning system that uses the training data to learn a translation model and a decoder that uses th translation model to translate new sentences.

2.3 Parallel Corpora

A text in a language and its translation in another language is called as parallel text. The first step of building a statistical machine translation system is compilation of a large collection of such bilingual text. In general such parallel corpora are not sentence-wise parallel and contain sentence insertions, deletions etc. One needs a further step, so called sentence alignment that extracts parallel translated sentences from this corpora. This step is needed as translation parameters and further statistics for word-alignment will be estimated from these sentence pairs. Some known parallel corpora are; Europarl corpus [5] from European Parliament proceedings for 11 languages, Hansards corpus from Canadian Hansards collection in English and French with 1.3 sentences and LDC corpus.[2] [3]

There are many different approaches for sentence alignment. Language independence is the common property of these different approaches. Brown et al. [19] used token/word counts with the assumption that sentences which are translation of each other should not differ wildly in the number of tokens. Gale and Church [20] calculated character length counts with a similar assumptions. Melamed [21] used word translation correspondence and Moore [22] presented a hybrid approach combining word translation correspondence and sentence length counts. Sentence-aligned parallel corpora is usually preprocessed by tokenization, filtering long sentences and

[2]Canadian Hansards Corpus is available at http://www.isi.edu/natural-language/download/hansard/

[3]LDC Corpus is available at http://www.ldc.upenn.edu/

lower-casing the sentences.

Obviously, for accurate calculation of statistics, one needs large amounts of training data. Koehn [5] gives some statistics about multilingual corpus collected in the Europarl project. This corpus contains about a million sentences for all languages which for some non-European language pairs such as Inuktitut, Hindi, Turkish may not be easy to obtain. This can be further complicated by the nature of the languages involved. In this case, researchers should preprocess parallel corporas and/or adapt translation systems to get the maximum gain.

2.4 The Translation Model

An SMT system estimates translation parameters from parallel corpora by statistical methods. Initial assumption of the translation system is that every Turkish (t) sentence is a possible (not necessarily correct) translation of every English e sentence with some translation probability. For every pair of sentences (e, t), $P(t \mid e)$ is the probability of generating target sentence $t{=}t_1, t_2, \ldots t_n$ for a given source sentence $e{=}e_1, e_2, \ldots e_m$.

Thus given some output sequence (e) one tries to find

$$t^* = \arg\max_t P(t|e) \tag{2.1}$$

as that input (Turkish) sentence that maximizes the probability of giving rise to the specific output (English) sentence e. Due to this approach, a source sentence have many acceptable candidate translations in the target language. For example, an English sentence e can be correctly translated into Turkish with many different sentences. So, given the (observed) sentence e, presumably the translation of an original sentence t, one tries to recover the most likely sentence t^* that could have given rise to e. Thus in a machine translation setting, e is the *source* language

sentence for which we seek the most likely *target* language sentence, t^*. There are two main approaches to model the posterior probability, $P(t \mid e)$; decomposing onto components and direct calculation.

2.4.1 Noisy-Channel Model

Most formulations of statistical machine translation views translation as a noisy-channel signal recovery process as shown in Figure 2.2.

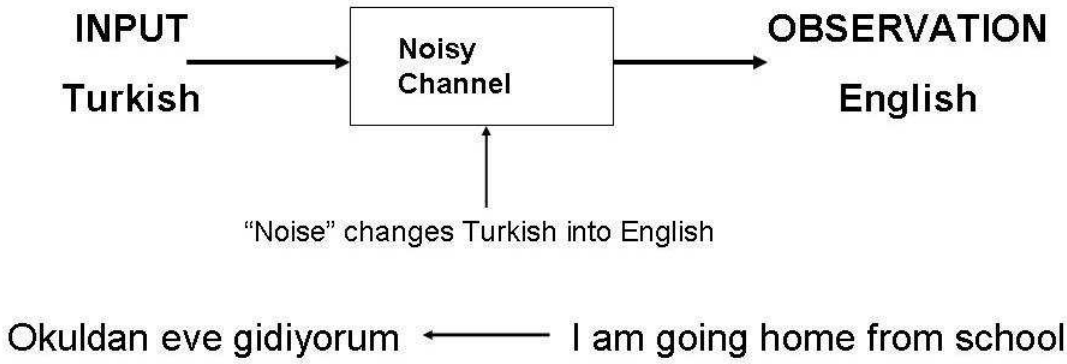


Figure 2.2: Noisy Channel

In noisy-channel model, one tries to recover the original form of a signal that has been *corrupted* as it is transmitted over a noisy channel. In this context, *corruption* corresponds to the translation of sentence $t{=}t_1, t_2, \ldots t_n$, into a sentence $e{=}e_1, e_2, \ldots e_m$ in a different language. By using Bayes' law;

$$t^* = \arg\max_t P(t \mid e) = \arg\max_t \frac{P(e \mid t)P(t)}{P(e)} = \arg\max_t P(e \mid t)P(t) \qquad (2.2)$$

since e is constant for all candidate sentences t. This formulation is known as *Fundamental Equation of Machine Translation*. This decomposition has two components which allow separate modelling of the adequacy (translation of words of the source sentence) and fluency (word order of target sentence).

The first component $P(e \mid t)$, called the *translation model*, gives the probability of translating t into e and models whether the words in English sentence are in general, translations of words in Turkish sentence. Given a pair of sentences e and t, it assigns probabilities $P(e \mid t)$ to possible sentences, t, given the source sentence e based on how good words or phrases in e are translated to words or phrases in t, that is, translation model assigns higher probabilities to sentences in which the words or phrases are *good* translation of words or phrases in the source sentence e. The translation model relies on model parameters that are estimated from sentence-aligned parallel texts [12]. These parameters include translation, distortion, and fertility probabilities. Translation model is learned by an iterative expectation maximization algorithm that aligns words and extracts translation probabilities.

The second component, $P(t)$, is the prior probability of target sentence and called as the *language model*. $P(t)$ models target (Turkish) sentences by assigning the sentence t, a certain probability among all possible sentences in the source language. In general, syntactically well-formed sentences will be assigned higher probabilities than ill-formed or *word-salad* sentences. Most recent statistical machine translation approaches rely on the language model to model target language sentences. It helps to avoid syntactically incorrect sentences.

Language model is based on the well-known n-gram counting and extraction of probabilities. A sentence t with a sequence of words $t{=}t_1, t_2, \ldots t_n$, language model $P(t)$ gives the probability of syntactic correctness of sentence t with the formulation;

$$P(t) = P(t_1 t_2 \ldots t_n) = P(t_1)P(t_2 \mid t_1)P(t_3\ t_1 t_2) \ldots P(t_n \mid t_1 \ldots t_{n-1}) \qquad (2.3)$$

For long sentence, it is not feasible to calculate the probability $P(t_n \mid t_1 \dots t_{n-1})$. Therefore, most approaches use an approximation to this probability by using a certain number of previous words in the calculations. The model using two previous words is called the trigram model

$$P(t_k \mid t_1 \dots t_{k-1}) \approx P(t_k \mid t_{k-2} t_{k-1}) \tag{2.4}$$

Similar to the translation model, the language model also requires large amount of data to estimate the probabilities. Even with large amount of data it is possible to face some unobserved word triples so that computation in 2.3 ends up being 0 . For such word sequences, it is preferred to assign a low probability instead of zero probability. N-gram smoothing (add-one, interpolation or backoff) is used to assign a low probability for such unseen n-grams.

The translation model is trained using the parallel corpora by determining the translations of individual tokens while language model is trained by a monolingual data of target language. The two models can be estimated independently.

Figure 2.3 shows the structure of the statistical machine translation prototype with noisy-channel model.

2.4.2 The Log-Linear Model

Another alternative for modelling the posterior probability $P(t \mid e)$ is the direct modelling with a log-linear approach [16, 23]. This approach is the generalized version of noisy-channel model which is used when the system is powered with extra features in addition to the language and translation models. Some typical features are phrase translation probabilities, lexical translation probabilities, reordering models and word penalty. This approach models $P(t \mid e)$ as a weighted combination of feature functions. Each feature such as language model, sentence-length model,

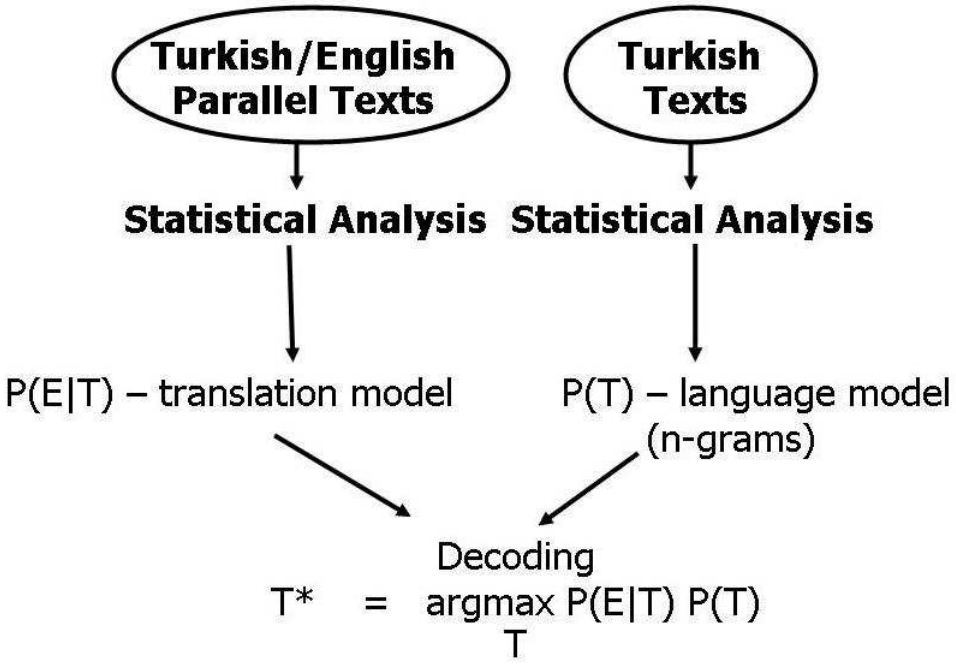


Figure 2.3: English to Turkish statistical machine translation structure with noisy-channel model

phrase-based translation model that effects the translation is expressed by a feature function and then the posterior probability is then the sum of these feature functions $f_i(t, e)$ with a model weight λ_i for $i = 1 \dots I$. The posterior probability is approximated by

$$P(t \mid e) = p_{\lambda_1^I}(t \mid e) = \frac{exp[\sum_1^I \lambda_i f_i(t, e)]}{\sum_{t'} exp[\sum_1^I \lambda_i f_i(t', e)]} \tag{2.5}$$

Similar to the noisy channel approach, since e is constant for all candidate t's, in the search problem, the renormalization introduced by divisor is eliminated.

$$t^* = \arg\max_t P(t \mid e) = \arg\max_t \sum_1^I \lambda_i f_i(t, e) \tag{2.6}$$

In log-linear approach, training process turns out to be an optimization problem of the model parameters. The best suitable weights are determined on a training data to maximize the performance of translation system. With a maximum entropy framework [24, 25]

$$\lambda_1^{I^*} = \arg\max_{\lambda_1^I} \sum_1^S \log p_{\lambda_1^I}(t_S|e_S) \tag{2.7}$$

where S is the number of sentences in the training data.

Optimizing model parameters does not always mean that these parameters are optimal with respect to the translation quality. Another alternative is minimum error rate training [26] that uses the n-best lists obtained with the current best weights and tries to find a better set of weights that reranks the n-best list to obtain a better score.

Figure 2.4 shows the structure of the statistical machine translation prototype with log linear model.

2.5 Translation Approaches

Many statistical machine translation systems use very similar training phases but they show differences in the definition of translation unit. SMT initially started with word-based models. After observing that word translation is context dependent and words tend to be translated as groups, phrase-based approaches have introduced *phrases* which in this context denote any sequence of tokens (that may or may not be linguistically meaningful). More recently, factored models use factors as translation unit that exploit richer linguistic information such as word roots, parts-of-speech and morphological information. Recently, there has been substantial work on including syntactic information in the translation process.

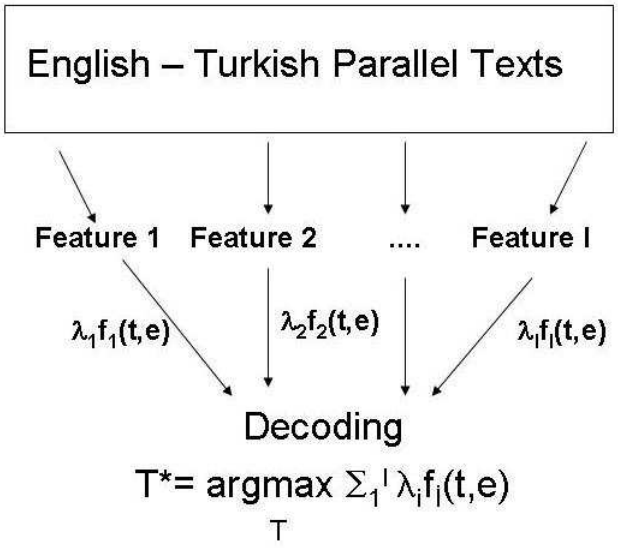


Figure 2.4: English to Turkish statistical machine translation structure with log linear model

2.5.1 Word-Based Approach

The initial work in statistical machine translation was started with IBM's Candide project [13]. IBM's word-based model [12] used a purely word-based approach without taking into account any of the morphological or syntactic properties of the languages.

IBM models are based on basically counting the source and target word occurrences and positions in the same sentence pairs over all possible alignments. A hidden variable, alignment $A{=}a_1, a_2, \ldots a_n$, is introduced to define all possible source and target word alignments. The translation probabilities and best alignmentare iteratively calculated over these alignments by expectation maximization algorithm.[4]

[4]Best alignment is also called as Viterbi alignment

$$P(e \mid t) = \sum_{A} P(e, a_i \mid t) \tag{2.8}$$

In the IBM models, there is only one restriction in the word alignments: a source word may translate into many target words but the reverse is not allowed. Figure 2.5 shows a two-sentence corpus with some possible word alignments and one illegal alignment. At the end of iterative training of this two-sentences corpus, the probabilities $P(house \mid ev)$ and $P(blue \mid mavi)$ will converge to 1 as word pair `blue` and `mavi` occurs in both of the sentences. However, there is not enough information to distinguish the translations of words `büyük` and `kitap` so the translation probabilities $P(big \mid büyük)$, $P(big \mid kitap)$, $P(book \mid büyük)$ and $P(book \mid kitap)$ will be almost same and close to 0.5.

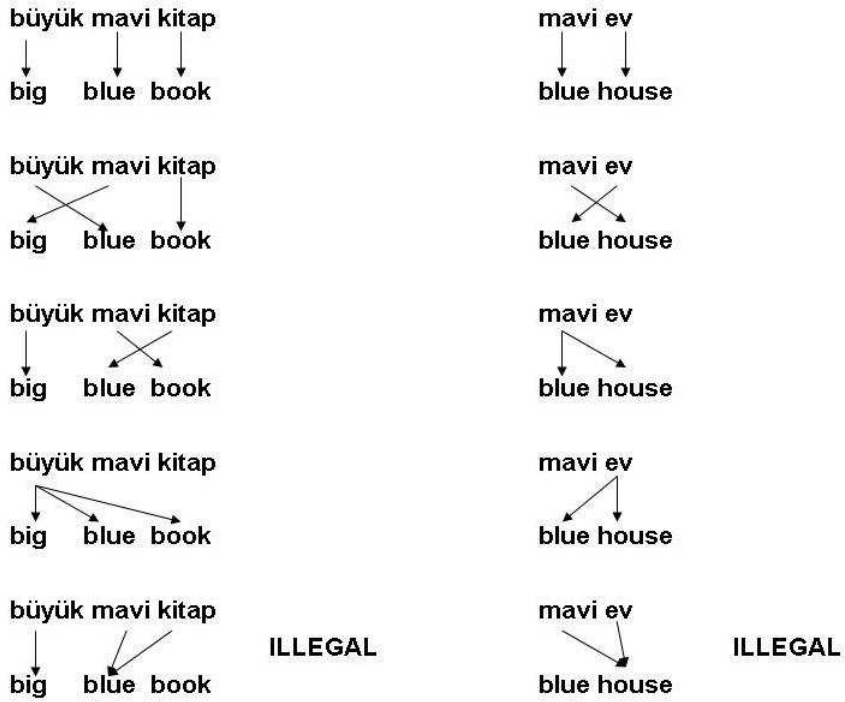


Figure 2.5: Some possible word alignments for a two-sentenced corpus

IBM introduced a five-stage approach to model $P(e \mid t)$ that iteratively learns the translation, distortion[5], fertility[6] and null translation [7] probabilities. IBM Model 1 just models the translation probabilities with the initial guess that all connections for each target position is equally likely without taking into consideration the order and location of the words. Model 2 models the distortion probability in addition to the translation probabilities. IBM Model 3 includes fertility probabilities, null generation probabilities and a reverse distortion probability in place of distortion probability. Model 4 models the same probabilities with Model 3 but using a more complicated reordering model and Model 5 fixes the deficiency. Later, Och and Ney showed that Model 6 [23] -the log-linear combination of Model 4 and HMM Model [27]- gives better results. They also implement the GIZA++ tool that is the most common used training tool for word alignments.

2.5.2 Phrase-Based Approach

The main shortcoming of the IBM models and so the word-based approaches is the one-to-many relationship between source and target words. As a result of this constraint, the word alignments that are learnt for the language pair does not reflect the real alignments and many words are left as unaligned if the languages have different fertilities. In English to Turkish word alignment, each word of a Turkish sentence may produce any number of English words (including zero word) but it is impossible to group any number of Turkish words to produce a single English word. Figure 2.6 shows a word-based alignment for the Turkish-English sentence pair `Yarın Kanada'ya uçacağım` and `Tomorrow I will fly to Canada`. In IBM models, as it is not allowed a source word to match more than one word; word `uçacağım` aligned only to the word `fly` and similar situation also occurs for the word `Kanada'ya`.

[5]distortion models how likely is it for a word t occurring at position i to translate into a word e occurring at position j, given target sentence length n and source sentence length m

[6]fertility models how likely is it to translate a word t into n words $e1e2e3 \ldots en$

[7]null translation models how likely is it for a word t to be spuriously generated

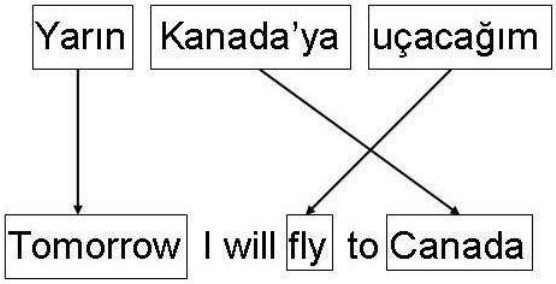


Figure 2.6: A word-based alignment

One other shortcoming of the word-based approaches is the lack of context information while translating. Generally, words tend to be translated in groups and word by word translation does not always give the actual meaning of a whole phrase. For any word, the translation and position in the target language may differ depending on the nearby words which is also called as localization effect. For example; the verb `quit` is translated as `bırakmak` in the context `quit smoking` and as `çıkmak` in the context `quit the program`. Word-based models only employ the language models for these cases which is not sufficient alone.

Such limitations of basic word-based models prompted researchers to exploit more powerful translation models that uses bilingual phrases. First, phrase-based approaches started with alignment templates [16] and continued with many others [14, 15, 17, 28]. Phrase-based models extract phrase translations allowing explicit modelling of context and some local word reorderings in translation.[8] Figure 2.7 shows a phrase alignment for the sentence pair above.

Basically, phrase translations are extracted from the combination of bi-directional word alignments which allows a many-to-many mapping. To extract the phrases that are consistent with word alignments, a combination of the intersection and union of

[8]Despite the linguistic meaning, a phrase in this context is defined as any contiguous sequence of words.

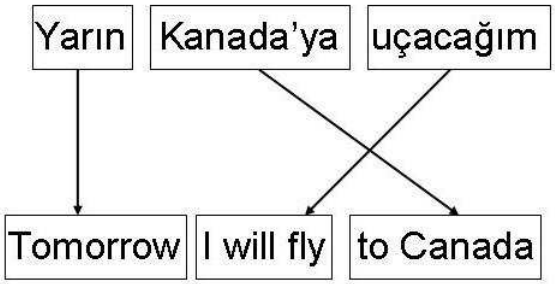


Figure 2.7: A phrase-based alignment

these word alignments are merged by some rules.[9]. It should be noted that phrases should be composed of continuous word sequences. Figure 2.8 shows an example of word mapping matrix and possible phrases.

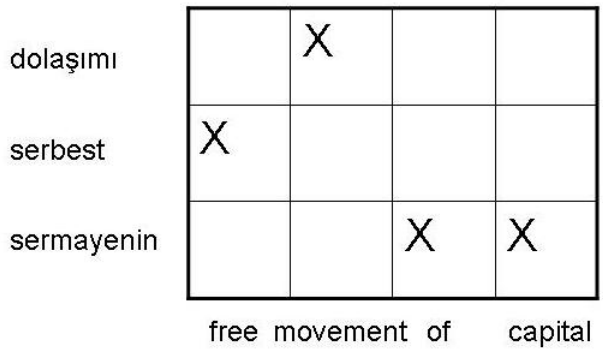
Phrase-based models introduce a phrase translation probability $\phi(\bar{e} \mid \bar{t})$, the probability of the translation of source phrase $\bar{e}$ given the target phrase $\bar{t}$, in place of word translation probability. Phrases that are common enough in the training data are obtained by the relative frequency

$$\phi(\bar{e} \mid \bar{t}) = \frac{count(\bar{t}, \bar{e})}{\sum_{\bar{e}} count(\bar{t}, \bar{e})} \tag{2.9}$$

A portion of a phrase table extracted from aligned Turkish-English parallel texts is shown in Table 2.1.

In phrase-based models, source sentence e is divided into I phrases as $e = ep_1, ep_2, \ldots ep_I$ with uniform probability distribution. Each of the source phrases ep_i are translated into target phrases tp_j to form the target sentence as $t = tp_1, tp_2, \ldots tp_J$. Although target phrases are reordered by a relative distortion probability distribution, generally most phrase translation models [15, 25] use weak reordering schemes

[9]For details http://www.isi.edu/licensed-sw/pharaoh/manual-v1.2.ps



(sermayenin,of), (sermayenin, of capital), (serbest,free) (serbest dolaşımı, free movement),(sermayenin serbest dolaşımı, free movement of capital)

Figure 2.8: A word matrix and possible phrases for a Turkish-English sentence pair

in order simplify the modelling. Some models [29,30] prefer a monotone translation where phrases are translated more or less in the order they appear in the source sentence. Clearly, this is a problem for language pairs with very different word orders. To overcome the monotonicity problem, Chiang [17] has introduced a hierarchical phrase-based model that can make longer distance reorderings.

Turkish phrase	**English phrase**	$\phi(t \mid e)$	$\phi(e \mid t)$
education , health and infrastructure	eğitim , sağlık ve altyapı	0.109	0.103
education , health and social	eğitim , sağlık ve sosyal	0.265	0.116
education , health and	eğitim , sağlık ve	0.299	0.121
education , health	eğitim , sağlık	0.369	0.136
education , poor health and	eğitim , yetersiz sağlık ve	0.014	0.002
education , poor health	eğitim , yetersiz sağlık	0.017	0.002
education , poor	eğitim , yetersiz	0.003	0.024

Table 2.1: A portion of the phrase table

2.5.3 Factor-Based Approach

Although phrase-based models improve upon word-based models, both approahes have a common shortcoming in surface representation of words. Basically, neither model integrates an explicit linguistic information into the translation model. Therefore words with morphological similarities are treated as separate tokens and unrelated. For example, the morphologically related Turkish words `faaliyet (activity)` and `faaliyetler (activities)` are treated as totally different words and occurrence of one does not give any information about the other word, although they share common roots and the second is the plural form of the first word. If in the training, the translation pair (`faaliyet,activity`) is learned and the system encounters the new word `activities`, the decoder will not be able to translate although the root is known by the translation model.

Very recently, the factored model approach that is an extension of the phrase-based models has been proposed to integrate some linguistic and lexical information such as root, features, pos information, morphology, etc. into the translation process [18]. Factored models aim to eliminate the data sparseness problem by translating the lemmas and morphological information separately instead of surface words. Figure 2.9 shows the general idea behind factored translation.[10]

Experiments show that factored models are suitable to languages with parallel inflectional morphology which usually happens to be mostly inflectional, such as German, Spanish and Czech but not preferable if the languages are very distant and richer morphology is on the target side. When translating into a complex morphology language from poor morphology language such as English to Turkish, although factored models can show a success for translating lemmas, poor morphological information of English fails to generate the morphemes in the Turkish side especially derivational morphemes. Turkish morphemes are mostly expressed in English by function words, prepositions, auxiliary verbs etc. Only very limited

[10]Figure is taken from site http://www.statmt.org/moses/?n=Moses.FactoredModels

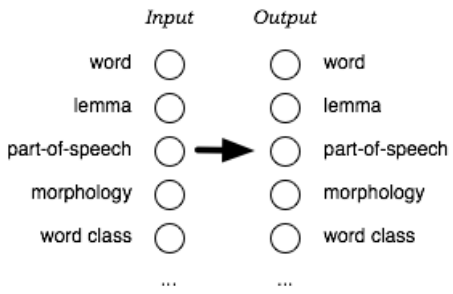


Figure 2.9: Factor-based Translation

morphological information can be translated from the source language English into Turkish. Additionally, the current synchronous modelling of factored models only allow translations within specific phrases but Turkish sometimes collect morpheme information for one surface word from different English phrases.

2.6 Decoding

Given a translation model and a new sentence, a decoder searches for a target sentence that maximizes equation 2.2.

Decoding tries to find the translation of this sentence by maximizing fundamental equation of statistical machine translation. Statistical translation decoders are responsible for the search process that is implied by the arg max of the equation. The decoder combines the evidence from $P(f \mid e)$ and $P(e)$ maximizing the product of two models in the noisy-channel model and sums the evidences from different models with different weights in log linear model to find the best translation.

Decoders take a source sentence and first segment it into all possible tokens. In a left to right fashion, tokens of source sentence (grouped into phrases if using the phrase-based approaches) are then translated and moved around into many

possible target language token sequences and scored with probabilities provided by the components of translation model. But the set of possible target sentences grow up exponentially hence the search process is controlled to reduce the search space by hypothesis re-combination and pruning heuristics.

As optimal decoding is known to be NP-complete [31], researchers have resorted to approximate algorithms that rely on certain heuristics. Greedy algorithms are used in first word-based decoders such as ISI Rewrite decoder [32, 33]. State-of-the-art algorithms are stack-based beam search algorithms and are used in phrase-based and factored-based decoders such as Pharaoh and Moses [34].

2.7 Automatic Evaluation of Translation

Evaluation is one of the most challenging problems in machine translation. Researchers developing new models are expected to evaluate the changes in performance by some means. To evaluate the performance of an SMT system, one should compare the decoded sentences with reference sentences and score them based on how grammatical they are and how accurately they reflect the source sentence. The best way of evaluating an MT system is ultimately based on human judgment with which, aspects of translation quality, such as adequacy, fidelity and fluency can be judged. On the other hand, human evaluation is however slow and labor intensive.

In evaluation, if a lot of words in the candidate translation occur in the reference translation, then the candidate is considered *adequate*, while if a lot of n-grams of words (especially for large n) occur in the reference, then the candidate is considered *fluent*. To analyze the systems quickly and also inexpensively, researchers need an automatic way of evaluation.

Initially, for automatic machine translation evaluation, metrics such as WER, PER, and mWER used in speech recognition are used. *WER (word error rate)* computes the number of substitutions, insertions and deletions among the decoded

sentence and references by using the edit distance. A lower WER indicates better translation. *PER (position-independent word error rate)* [35] is very similar to WER metric but ignores word order. A sentence is treated as a bag-of-words as an expectation of a perfect word order is usually too strict, especially for flexible word order languages. *mWER (multi-reference word error rate)* [36] is very similar to PER and is used for systems with multiple reference sentences. All these metrics were originally developed for speech recognition evaluation and just evaluate adequacy as it is sufficient for speech evaluation, as word order does not play an important role.

Later, new metrics such as, *NIST* [37], *BLEU* [7] and *METEOR* [38] incorporated fluency into machine translation evaluation. This group of metrics use n-gram co-occurrences to find similarity of the candidate translation and the reference sentence/s. BLEU uses modified precision by calculating geometric mean of n-grams (general usage n up to 4), NIST is variant of the BLEU metric and uses the weighted precision of matching n-grams (give weights depending on n-gram frequencies), METEOR is similar to BLEU, tries to fix some of deficiencies of BLEU. METEOR uses the harmonic mean of 1-gram precision and incorporate recall, and additionally checks stems and WordNet [39] relations for the synonyms for the words that do not match in the reference sentences. As shorter sentences tend to have higher scores, all these metrics use a factor that penalizes the short sentences.

2.7.1 BLEU in detail

BLEU is the most popular measure that has been proposed and used as an automatic way of gauging MT quality. BLEU scores the output of an MT system by comparing each sentence to a set of reference translations using n-gram overlaps of word sequences. The standard BLEU computation is;

$$BLEU = BP \cdot exp[\sum_{n=1}^{N} w_n \log p_n] \tag{2.10}$$

where BP is the brevity penalty to penalize the long candidate translations, p_n is the modified precision and w_n is the weight for n-grams (uniform, most commonly $N = 4$).

Chapter 3

ENGLISH TO TURKISH STATISTICAL MACHINE TRANSLATION

3.1 Challenges

Statistical machine translation poses many lexical and structural challenges such as word sense ambiguities, lexical gaps between languages, word and constitutient order differences, translation of idioms, treatment of out-of-vocabulary words and more. In English-to-Turkish statistical machine translation, two of the above problems comprise the main motivation points of this thesis. Firstly, English and Turkish are rather distant languages, with different word orders that result in a huge lexical gap between the languages. Furthermore the English-Turkish available parallel corpus is very limited compared to other language pairs that have been extensively studuied.[1]

[1]Europarl [5] parallel corpus for English-German and English-French pairs have over 1 million sentences.

3.1.1 Turkish Morphology

Turkish is an Ural-Altaic language, having agglutinative word structures with productive inflectional and derivational processes. Turkish word forms consist of morphemes concatenated to a root morpheme or to other morphemes, much like *beads on a string*. Except for a very few exceptional cases, the surface realizations of the morphemes are conditioned by various regular morphophonemic processes such as vowel harmony, consonant assimilation and elisions. Further, most morphemes have phrasal scopes: although they attach to a particular stem, their syntactic roles extend beyond the stems. The morphotactics of word forms can be quite complex when multiple derivations are involved. For instance, the derived modifier `sağlamlaştırdığımızdaki` can be translated into English literally as `(the thing existing) at the time we caused (something) to become strong`. Obviously this word is not a word that one would use everyday. Turkish words (excluding non-inflecting frequent words such as conjunctions, clitics, etc.) found in typical running text average about 10 letters in length. The average number of bound morphemes in such words is about 2. The word `sağlamlaştırdığımızdaki` would be broken into surface morphemes as follows:

```
sağlam+laş+tır+dığ+ımız+da+ki
```

Starting from an adjectival root `sağlam`, this word form first derives a verbal stem `sağlamlaş`, meaning `to become strong`. A second suffix, the causative surface morpheme `+tır` which we treat as a verbal derivation, forms yet another verbal stem meaning `to cause to become strong` or `to make strong (fortify)`. The immediately following participle suffix `+dığ`, produces a participial nominal, which inflects in the normal pattern for nouns (here, for 1^{st} person plural possessor which marks agreement with the subject of the verb, and locative case). The final suffix, `+ki`, is a relativizer, producing a word which functions as a modifier in a sentence, modifying a noun somewhere to the right.

However, if one further abstracts from the morphophonological processes involved one could get a lexical form

```
sağlam+lAş+DHr+DHk+HmHz+DA+ki
```

In this representation, the lexical morphemes except the lexical root utilize metasymbols that stand for a set of graphemes which are selected on the surface by a series of morphographemic processes which are rooted in morphophonological processes some of which are discussed below, but have nothing whatsoever with any of the syntactic and semantic relationship that word is involved in. For instance, `A` stands for back and unrounded vowels `a` and `e`, in orthography, `H` stands for high vowels `ı`, `i`, `u` and `ü`, and `D` stands for `d` and `t`, representing alveolar consonants. Thus, a lexical morpheme represented as `+DHr` actually represents 8 possible allomorphs, which appear as one of `+dır, +dir, +dur, +dür, +tır, +tir, +tur, +tür` depending on the local morphophonemic context.

The productive morphology of Turkish implies potentially a very large vocabulary size: noun roots have about 100 inflected form and verbs have much more [40]. These numbers are much higher when derivations are considered; one can generate thousands of words from a single root when, say, only at most two derivations are allowed. For example, a recent 125M word Turkish corpus that we have collected has about 1.5 M distinct word forms. This is almost the same number of distinct word forms in the English Gigaword Corpus which is about 15 times larger.

3.1.2 Contrastive Analysis

Turkish and English have many differences that make the English-to-Turkish machine translation a challenging issue:

1. Typologically English and Turkish are rather distant languages in certain basic linguistic dimensions: Watkins provides a summary of language typologies

where English and Turkish fall in different categories with respect to word order.[2] While English has very limited morphology with a rather rigid subject-verb-object constituent order, Turkish is an agglutinative language with a very rich and productive derivational and inflectional morphology, and a very flexible (but subject-object-verb dominant) constituent order. Barber [41] states that according to word formation English is an analytic language while Turkish is a synthetic language with lots of morphemes attached to a free root morpheme. In Turkish, it is possible to form 24 acceptable sentences from a 4-word string. Below some possible Turkish sentences are shown for the sentence $\texttt{Yesterday}_1$, $\texttt{Ali}_2$ $\texttt{saw}_3$ $\texttt{his}_4$ $\texttt{new}_5$ $\texttt{friend}_6$, that can be used in distinct discourse contexts.

$\texttt{Dün}_1$ $\texttt{Ali}_2$ $\texttt{yeni}_5$ $\texttt{arkadaşını}_{4,6}$ $\texttt{gördü}_3$
Ali dün yeni arkadaşını gördü
Ali dün gördü yeni arkadaşını
Ali gördü yeni arkadaşını dün
Gördü Ali yeni arkadaşını dün
Gördü dün Ali yeni arkadaşını
Yeni arkadaşını dün gördü Ali
Dün yeni arkadaşını gördü Ali

2. Turkish verbs can have two types of suffixes: personal and tense suffixes, and optionally can carry a variety of others. In English only tense suffixes are attached to the verbs, the rest is expressed separately, which causes a Turkish verb map to an English verb phrase. Some Turkish verbs and English counterpart verb phrases is shown below.

($\texttt{içer}_1\texttt{mez}_2$, $\texttt{does}_2$ $\texttt{not}_2$ $\texttt{contain}_1$)
($\texttt{yürüt}_1\texttt{ül}_2\texttt{ecek}_3\texttt{tir}_4$, $\texttt{will}_3$ $\texttt{be}_4$ $\texttt{continue}_1\texttt{d}_2$)
($\texttt{gör}_1\texttt{em}_2\texttt{iyor}_3\texttt{du}_4\texttt{m}_5$, $\texttt{I}_5$ $\texttt{was}_4$ $\texttt{un}_2\texttt{able}_3$ $\texttt{to see}_1$)

[2] http://www.sjsu.edu/faculty/watkins/langtyp.htm

3. As Turkish verbs carry person suffixes, the subject pronoun can be deleted most of the time. In English, pronouns are always a part of the sentence. Some Turkish sentences with deleted pronouns in parenthesis and their English translations are shown below.

   ((Ben)$_1$ Okul$_2$a$_3$ git$_4$ti$_5$m$_6$, I$_{1,6}$ went$_{4,5}$ to$_3$ shool$_2$)
   ((Biz)$_1$ (sizin)$_2$ ev$_3$iniz$_4$e$_5$ gel$_6$di$_7$k$_8$, We$_{1,8}$ came$_{6,7}$ to$_5$ your$_{2,4}$ house$_3$)

4. In Turkish noun phrases, noun head is always placed at the end. In English noun phrases, noun head can take both pre-nominal and post-nominal modifiers.

   geçen$_1$ hafta$_2$ aldığı$_3$ yeşil$_4$ araba$_5$
   the green$_4$ car$_5$ that$_3$ he$_3$ bought$_3$ last$_1$ week$_2$

5. Inserting one sentence into another to make a more complex sentence is called embedding. In Turkish, sentences are embedded by concatenating suffixes or suffixes plus functional words to the verb. On the other hand, English embedded sentence preserves most of its constituents. Embedding done just by functional words such as that, who, which, etc. Some examples are;

   Herkes Ali'nin daha iyi bir yaşamı hakettiğini söylüyor
   Everybody says that Ali deserved a better life

   Japonya'da üç yıl yaşayan arkadaşım
   My friend who has lived in Japan for three years

   Ahmet kendisinin geleceğini söyledi
   Ahmet said that he would come

3.1.3 Available Data

The first step of building an SMT system is the compilation of a large amount of parallel text for accurate estimation of parameters. This turns out to be a significant problem for the Turkish and English pair because of the lack of such texts. We collected a less homogeneous corpus as there are not many and consistent sources for Turkish-English parallel texts. The only sources that we could find and access are, EU/NATO Documents, Foreign Ministry Documents, International Agreements, etc. In terms of news, the Balkan Times news paper produces some parallel Turkish - English text, but the Turkish side (at least) has enough typos and unnecessary word breaks to render it unusable without extensive work.

Although we have collected about many parallel texts, most of these require significant clean-up (from HTML/PDF sources). We cleaned about 60.000 sentences of these parallel texts. We used the subset of these sentences of 40 words/tokens or less as our training data, in order not to exceed the maximum number of words recommended for training the translation model.[3]

Dictionaries

Dictionaries and similar resources comprise an additional resource that bootstrap training of statistical alignment models and cover vocabulary that does not occur in the training corpus for obtaining more accurate alignments. Dictionaries provide possible correct word translation pair biases to the expectation maximization algorithm used in generating word-level alignments and increase translation probabilities that will help to obtain better alignments. Conventional dictionaries such as Harper-Collins Robert French Dictionary have been used as an additional source for the French-English translation developed by IBM [42].

Another interesting resource that can be used to help alignment, in place of

[3]Details of the corpus is in Chapter 3.4

a dictionary, is WordNet [43], a hierarchical network of lexical relations (such as synonyms) that words in a language are involved in. The Turkish WordNet [44] was built earlier, and is actually linked to the English WordNet using interlingual indexes, so that words in Turkish are indirectly linked to words in English that describe the same concept via these indexes. For example the synset `(toplamak, biriktirmek)` is linked with the English synset `(roll up, collect, accumulate, pile up, amass, compile, hoard)`. We generate a parallel data from these relations and integrate 12002 sentences into the training set.

3.2 Integrating Morphology

If one computes a word-level alignment between the components of parallel Turkish and English sentences one obtains an alignment like the one shown in Figure 3.1, where we can easily see Turkish words may actually correspond to whole phrases in the English sentence.

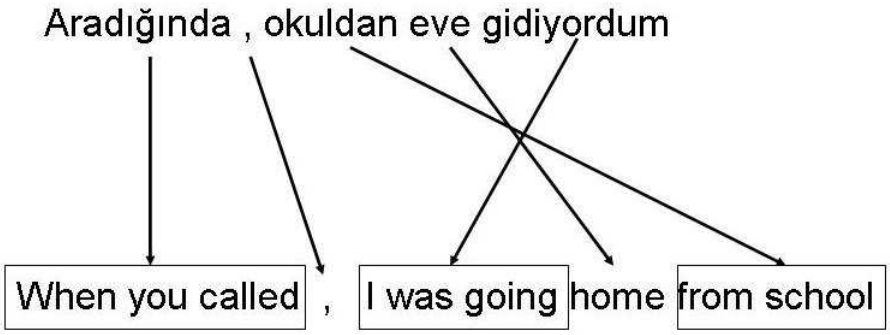


Figure 3.1: Word level alignment between a Turkish and an English sentence

A major problem with the word-based statistical machine translation systems

is that each word form is treated as a separate token and no explicit relationship between other words are defined. Because of this construct, any form of a word that is not in the training data (called as out-of-vocabulary words (OOV)) can not be translated. In the English-Turkish parallel corpora, it is very frequent to get a situation in which when even a word occurs many times in English part, the actual Turkish equivalent could be either missing or occur with a very low frequency, but many other inflected variants of the form could be present. As the productive morphology of Turkish implies potentially a very large vocabulary size, sparseness is an important issue given that we have very modest parallel resources available.

For example, Table 3.1 shows the inflected and derived forms of the root word `faaliyet (activity)` in the parallel texts we experimented with. Although the root appears many times, inflected and derived forms seems to appear rarely.

Therefore, if one considers each Turkish word as a separate token none of the forms in the corpus could help to learn other forms. This would be worse when very low frequency tokens would be removed from statistics as is typically done in language modeling, meaning that, most variants of words would possibly be dropped and language modeling would resort to out-of-vocabulary word smoothing processes that makes their statistics very unreliable.

Furthermore, if one wants to translate the phrase `in our activities`, decoder will not be able to produce the right word `faaliyetlerimizde` as there is no information about this word in the training set.

Consequently, initial exploration into developing a statistical machine translation system from English to Turkish pointed out that using standard models to determine the correct target translation was probably not a good idea. In the context of the agglutinative languages similar to Turkish (agglutinative language, similar morphological structure with Turkish), there has been some recent work on translating from and to Finnish with millions of sentences in the Europarl corpus [5]. Although the BLEU [7] score from Finnish to English is 21.8, the score in the reverse direction

Wordform	Count	Gloss
faaliyet	125	'activity'
faaliyetleri	89	'their activities'
faaliyetlerinin	44	'of their activities'
faaliyetler	42	'activities'
faaliyetlerini	41	'their activities (accusative)'
faaliyetlerin	28	'of the activities'
faaliyetlerde	16	'in the activities'
faaliyetlerinde	12	'in their activities'
faaliyetinde	10	'in its activity'
faaliyetlerinden	8	'of their activities'
faaliyetleriyle	5	'with their activities'
faaliyetlerle	3	'with the activities'
faaliyetini	2	'the activity (accusative)'
faaliyetteki	1	'that which is in activity/active'
faaliyetlerimiz	1	'our activities'
Total	427	

Table 3.1: Occurrences of forms of the word *faaliyet* 'activity'

is reported as 13.0 which is one of the lowest scores in 11 European languages scores. Also, reported *from* and *to* translation scores for Finnish are the lowest on average, even with the large number of sentences available. These may hint at the fact that standard alignment models may be poorly equipped to deal with translation from a poor morphology language like English to an complex morphology language like Finnish or Turkish.

The main aspect that would have to be seriously considered first is the Turkish productive inflectional and derivational morphology in English to Turkish statistical machine translation. A word-by-word alignment between an English-Turkish sentence pair has some Turkish words aligned to whole phrases in the English side. Certain English functional words are translated as various morphemes embedded into Turkish words. This shows us that for an accurate word alignment, we need to consider sublexical structures. For instance, the Turkish word `tatlandırabileceksek` could be translated as (and hence would have to be aligned to something equivalent to) `if we were going to be able to make [something] acquire flavor`.

This word could be aligned as follows (shown with co-indexation of Turkish surface morphemes and English words):[4]

(tat)$_1$(lan)$_2$(dır)$_3$(abil)$_4$(ecek)$_5$(se)$_6$(k)$_7$
(if)$_6$(we are)$_7$(going to)$_5$(be able)$_4$(to make)$_3$[something]
(acquire)$_2$(flavor)$_1$

The details of the model have to at least take into consideration a probabilistic model of the morpheme morphotactics in addition to models of higher level word order. This will certainly require certain non-trivial amendments to the translation models developed so far for various other language pairs. To overcome this problem, we decided to perform morphological analysis of both the Turkish and the English texts to be able to uncover relationships between root words, suffixes and function words while aligning them. As Turkish employs about 150 distinct suffixes, when morphemes are used as the units in the parallel texts, the sparseness problem can be alleviated to some extent. Thus for instance the word `faaliyetleriyle` was segmented into `faaliyet +ler +i +yle` and the English word `activities` was segmented as `activity+s`. We then observed that we could achieve a further normalization on the Turkish representation and improve statistics by using lexical morphemes discussed earlier. Figure 3.2 shows the morpheme alignment of Figure 3.1.

Table 3.2 shows the translation probabilities for some of the English function words and affixes with some Turkish function words and suffixes in the morphemic representation. It can be seen that the top alignment for most cases is usually the most likely one. Of particular interest is the alignment of `will` to the Turkish future tense marker lexical morpheme `+yAcAk`[5] which is usually surrounded by other morphemes marking other relevant morphological features when it appears in a verb.

[4]Note that on the English side, the filler for [something] would come in the middle of this phrase.

[5]This morpheme has 4 allomorphs that differ in the selection of the vowels and the elision of the initial consonant depending on the morphographemic context.

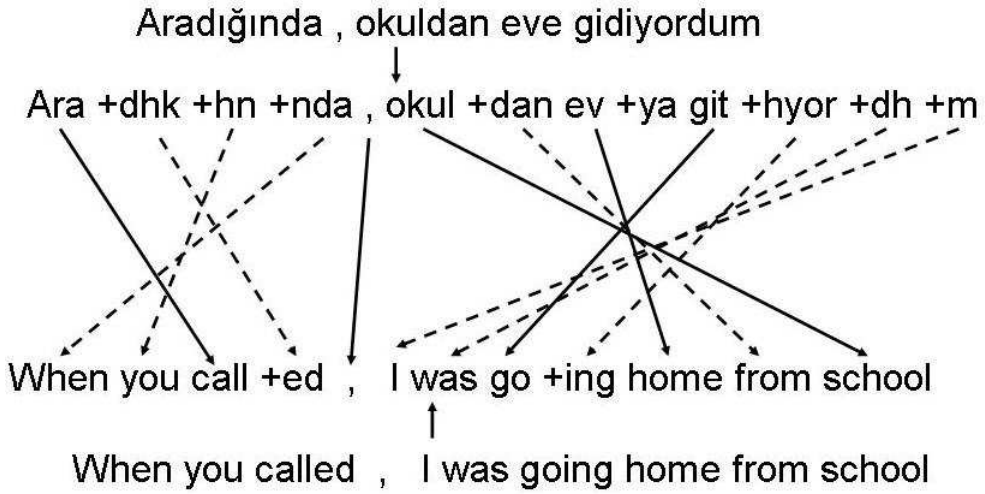


Figure 3.2: Morpheme level alignment between a Turkish and an English sentence

Also of interest are alignments of `should` to `+mAlH`, the Turkish necessitative mood marker, of `while` to `+yken`, the adverbial derivation suffix with the semantics `while`.

3.2.1 Related Work

Limitations of basic models in solving translation problems of language pairs with different morphological complexities prompted researchers to exploit morphological and/or syntactic/phrasal structure to increase the quality of parameters for the translation model and also to rely on smaller parallel texts. [14–16].

Niessen and Ney [45] use morphological decomposition to improve alignment quality. Yang and Kirchhoff [46] use phrase-based backoff models to translate words that are unknown to the decoder by morphologically decomposing the unknown source word. Corston-Oliver and Gamon [47] normalize inflectional morphology by stemming the word for German-English word alignment. Lee [48] uses a morphologically analyzed and tagged parallel corpus for Arabic-English statistical machine

e	**t**	$\phi(e\|t)$	$\phi(t\|e)$
has/have	+dhr +mhs	0.33	0.08
has/have	+mhs bulun +makta +dhr	0.4	0.05
has/have	sahip +dhr	0.72	0.06
+s	+lar +sh	0.80	0.89
+s	+lar	0.95	0.89
+s	+larh	0.86	0.56
+ed	+dh	0.79	0.52
+ed	+dhr +dh	0.6	0.26
+ed	+mhs	0.48	0.18
+ing	+hl +ma	0.40	0.06
+ing	+hl +mak	0.30	0.01
+ing	+hyor	0.19	0.11
will be	+dhr +hl +yacak	0.8	0.01
will be	+hl +yacak +dhr	0.83	0.02
will be	+hn +yacak +dhr	0.77	0.02
will have to	+ma +sh gerek +lh +dhr	0.12	0.01
will have to	+ma +sh gerek +yacak +dhr	0.25	0.01
will	+dhr +yacak	0.85	0.18
will	+yacak +dhr	0.70	0.18
will	+yacak	0.24	0.32
should be	+hl +malh +dhr	0.66	0.02
should be	ol +ma +sh gerek +hr	0.5	0.01
should be	ol +malh +dhr	0.5	0.01
should	+dhr +malh +dhr	0.85	0.10
should	+ma +lh	0.66	0.01
should	+malh +dhr	0.60	0.14
while	+hr +yken	0.62	0.09
while	+r +yken	0.77	0.09
while	+yken	0.37	0.16

Table 3.2: Alignments for various Turkish morphemes and English suffixes and function words

translation. Zolmann et al. [49] also exploit morphology in Arabic-English statistical machine translation. Popovic and Ney [50] investigate improving translation quality from inflected languages by using stems, suffixes and part-of-speech tags. Goldwater and McClosky [51] use morphological analysis on Czech text to get improvements in Czech to English statistical machine translation. Recently, Minkov et al. [52] used morphological post processing *on the output side* using structural information and information from the source side, to improve translation quality.

3.3 Pre-processing

Sparse data problem is a common challenge for most of the statistical machine translation systems. A big portion of the words is just seen only once in the corpus and for such words, it is not possible to obtain the translation probabilities robustly. The sparseness problem gets worse for the languages pairs such as English-Turkish as there is a huge morphological gab between the languages.

Moreover, for accurate estimation of parameters, one needs large amounts of data which for English-Turkish language pairs may not be easy to obtain with no substantial improvement expected in the near future. This can be further complicated by the nature of the languages involved as in our case. Thus we have to exploit our available resources maximally instead of relying on future availability of more data.

Our approach to solve the sparseness problem is to represent Turkish and English (to some extent) words with their morphological segmentation. We have used a morphological preprocessing to identify morphemes on both the Turkish and the English words to alleviate the data sparseness problem but more importantly to uncover relationships between the morphemes on the Turkish side with morphemes and function words on the English.

We use lexical morphemes instead of surface morphemes, as most surface distinc-

tions are manifestations of word-internal phenomena such as vowel harmony, and morphotactics. At the morpheme level, we have split the Turkish words into their lexical morphemes while English words with overt morphemes have been stemmed, and such morphemes have been marked with a tag.

3.3.1 Turkish

We segment the words in our Turkish corpus into lexical morphemes whereby differences in the surface representations of morphemes due to word-internal phenomena are abstracted out to improve statistics during alignment. The reason for using lexical morphemes is the allomorphs which differ because of local word-internal morphographemic and morphotactical constraints. Allomorphs almost always correspond to the same set of words/tags in English when translated. When surface morphemes are considered by themselves as the units in alignment, statistics get fragmented and the model quality drops. For example, to give the plural information within a word, Turkish has two different surface morphemes, `+ler` and `+lar`, both translated to `+s` in English side.

However, with lexical morpheme representation, we can abstract away such word-internal details and conflate statistics for seemingly different suffixes, as at this level of representation words that look very different on the surface look very similar. Employing this representation on the Turkish side and conflating the statistics of the allomorphs allowed us to improve the alignments. For instance, although the words `evinde (in his house)` and `masasında (on his table)` look quite different, the lexical morphemes except for the root are the same: `ev+sh+nda` vs. `masa+sh+nda`.

We should however note that although employing a morpheme based representations dramatically reduces the vocabulary size on the Turkish side, it also runs the risk of overloading the decoder mechanisms to account for both word-internal morpheme sequencing and sentence level word ordering.

As with many similar languages, the segmentation of a surface word is generally ambiguous. We first generate a representation using our morphological analyzer [53] that contains both the lexical segments and the morphological features encoded for all possible segmentations and interpretations of the word. For the sentence *gözden geçirilmiş katılım ortaklığı belgesi hakkında basın açıklaması*, the output of the morphological analyzer is shown in Table 3.3.

Then we perform morphological disambiguation using morphological features [54]. Table 3.4 shows the result of disambiguation for the above sentence.

Once the contextually salient morphological interpretation is selected, we replace the features with the lexical morphemes making up a word. The original Turkish sentence and our representation is shown below:

Original Sentence: `gözden geçirilmiş katılım ortaklığı belgesi hakkında basın açıklaması`
Morpheme Rep.: `göz+dan geç+hr+hl+mhs katılım ortaklık+sh belge+sh hak+sh+nda basın açıkla+ma+sh`

3.3.2 English

Similarly, we segment the words in our English corpus into part-of-speech tags to obtain a similar representation with Turkish morphemes and combine some morpheme statistics for auxiliary verbs `have` and `be`.

The English text was tagged using TreeTagger [55], which provides a lemma and a part-of-speech for each word. The tag set of TreeTagger tagset is an expanded version of Penn Treebank tagset [56].[6] For the verbs `be` and `have`, the second letter is specified as B and H, respectively. We drop the lemmas and just leave the tags for the verbs that are specified with B and H such as `(have,has,having,...)`, `(were,`

[6]Penn Treebank tags are listed in Appendix 1

Word	**Analysis**
gözden	göz +Noun+A3sg+Pnon(+DAn)+Abl
gözden	gözde +Adj^DB+Noun+Zero+A3sg(+Hn)+P2sg+Nom
geçirilmiş	geç +Verb(+Hr)^DB+Verb+Caus(+Hl)^DB+Verb+Pass+Pos(+mHS) +Narr+A3sg
geçirilmiş	geç +Verb(+Hr)^DB+Verb+Caus(+Hl)^DB+Verb+Pass+Pos(+mHS) +Narr^DB+Adj+Zero
katılım	katılım +Noun+A3sg+Pnon+Nom
ortaklığı	ortak +Noun(+lHk)+A3sg+Pnon+Nom^DB+Noun+Ness+A3sg(+sH) +P3sg+Nom
ortaklığı	ortak +Noun(+lHk)+A3sg+Pnon+Nom^DB+Noun+Ness+A3sg +Pnon(+yH)+Acc
ortaklığı	ortak +Adj(+lHk)^DB+Noun+Ness+A3sg(+sH)+P3sg+Nom
ortaklığı	ortak +Adj(+lHk)^DB+Noun+Ness+A3sg+Pnon(+yH)+Acc
ortaklığı	ortaklık +Noun+A3sg(+sH)+P3sg+Nom
ortaklığı	ortaklık +Noun+A3sg+Pnon(+yH)+Acc
belgesi	belge +Noun+A3sg(+sH)+P3sg+Nom
belgesi	belge +Noun(+ZH)^DB+Adj+Almost
hakkında	hak +Noun+A3sg(+Hn)+P2sg(+DA)+Loc
hakkında	hak +Noun+A3sg(+sH)+P3sg(+ndA)+Loc
basın	bas +Noun+A3sg(+Hn)+P2sg+Nom
basın	bas +Noun+A3sg+Pnon(+nHn)+Gen
basın	bas +Verb+Pos+Imp(+yHn)+A2pl
basın	basın +Noun+A3sg+Pnon+Nom
açıklaması	açıkla +Verb+Pos(+mA)^DB+Noun+Inf2+A3sg(+sH)+P3sg+Nom
açıklaması	açıkla +Verb+Pos(+mA)^DB+Noun+Inf2(+ZH)^DB+Adj+Almost

Table 3.3: Morphological Analyzer output

gözden göz+Noun+A3sg+Pnon+Abl
geçirilmiş geç+Verb^DB+Verb+Caus^DB+Verb+Pass+Pos+Narr+A3sg
katılım katılım+Noun+A3sg+Pnon+Nom
ortaklığı ortaklık+Noun+A3sg+P3sg+Nom
belgesi belge+Noun+A3sg+P3sg+Nom
hakkında hak+Noun+A3sg+P3sg+Loc
basın basın+Noun+A3sg+Pnon+Nom
açıklaması açıkla+Verb+Pos^DB+Noun+Inf2+A3sg+P3sg+Nom

Table 3.4: A sample disambiguator output

`was, been, being,...)`and `(are,is,am)` where the tags give enough information alone.

For the sentence `the achievement of the colleagues whom I named just now and others has been outstanding here`, the output of the tagger is shown in Table 3.5.

We augmented the TreeTagger output with some additional processing for handling derivational morphology. We dropped any tags which did not imply an explicit morpheme or an exceptional form. For instance, the English word `colleagues` is segmented as `colleague +NNS` with its tag but the word `achievement` is represented by removing `+NN` as its original form. Table 3.6 provides the subset of the tags that we used in our examples for the sake of being self-contained.

To make the representation of the Turkish texts and English texts similar, tags are marked with a '+' at the beginning of all tags to indicate that such tokens are treated as *morphemes*. The original sentence and our representation is shown below:

Original Sentence: `the achievement of the colleagues whom I named just now and others has been outstanding here`
Morpheme Rep.: `the achievement of the colleague+NNS whom I name+VVD just now and other+NNS +VHZ +VBN outstanding here`

Word	**Part-Of-Speech**	**Lemma**
The	DT	the
achievement	NN	achievement
of	IN	of
the	DT	the
colleagues	NNS	colleague
whom	WP	whom
I	PP	I
named	VVD	name
just	RB	just
now	RB	now
and	CC	and
others	NNS	other
has	VHZ	have
been	VBN	be
outstanding	JJ	outstanding
here	RB	here

Table 3.5: TreeTagger output

Part-of-Speech	**Tags**
Noun, Plural	NNS
Verb, Base form	VB, VH
Verb, Past Tense	VVD, VBD, VHD
Verb, Gerund or present participle	VVG, VBG, VHG
Verb, Past Participle	VVN, VBN, VHN
Verb, 3rd person singular present	VVZ, VBZ, VHZ
Verb, Non-3rd person singular present	VVP, VBP, VHP

Table 3.6: Subset of tags used in English sentences

3.3.3 A baseline representation

A typical sentence pair and its representation in our baseline data looks like the following

> **Turkish:** `katılma ortaklığının uygulanması, ortaklık anlaşması çerçevesinde izlenecektir`
> **Baseline Rep.:** `kat+hl+ma ortaklık+sh+nhn uygula+hn+ma+sh , ortaklık anlaşma+sh çerçeve+sh+nda izle+hn+yacak+dhr`
> **English:** `the implementation of the accession partnership will be monitored in the framework of the association agreement`
> **Baseline Rep.:** `the implementation of the accession partnership will be monitor+vvn in the framework of the association agreement`

3.3.4 Content Words

The localization of the content words is an important issue in the morphemic representation. The translation should be content-word oriented where the translation of content words should be completed before the placement of morphemes. We do not actually try to determine exactly which morphemes are actually translated but rather determine the content words and then associate translated morphemes and functional words with the right content word depending on the words in the neighborhood. The resulting sequence of root words and their bags-of-morpheme can be run through a morphological generator which can handle all the word-internal phenomena such as proper morpheme ordering, filling in morphemes or even ignoring spurious morphemes, handling local morphographemic phenomena such as vowel harmony, etc. We propose to use the content training set as a start point.

From the morphologically segmented corpora we also extract for each sentence

the sequence of roots for open class content words (nouns, adjectives, adverbs, and verbs). For Turkish, this corresponds to removing all morphemes and any roots for closed classes. For English, this corresponds to removing all words tagged as closed class words along with the tags such as `+VVG` above that signal a morpheme on an open class content word. We use this to augment the training corpus and bias content word alignments, with the hope that such roots may get a better chance to align without any additional *noise* from morphemes and other function words.

A typical sentence pair of content words looks like the following;

Turkish: `kat ortaklık uygula ortaklık anlaşma çerçeve izle`
English: `implementation accession partnership monitor framework association agreement`

3.4 Corpus Statistics

Table 3.7 presents various statistical information about the train and test data that we used during the research. The sentences were sentence aligned using Microsoft Research Bilingual Sentence Alignment Tool[7].

One can note that there is a difference between the number of sentences in the basic training set and the content word training set. This is because the training set in the first row of 3.7 was limited to sentences on the Turkish side which had at most 90 tokens (roots and bound morphemes) in total in order to comply with requirements of the GIZA++ alignment tool. However, when only the content words are included, we have more sentences to include since much less number of sentences violate the length restriction when morphemes/function word are removed. For language models in decoding and n-best list rescoring, we use, in addition to the training data, a monolingual Turkish text of about 100,000 sentences in a segmented

[7]available at http://research.microsoft.com/ ~bobmoore/

Table 3.7: Statistics on Turkish and English training and test data,and Turkish morphological structure

TURKISH	**Sent.**	**Words (UNK)**	**Unique Words**
Train	45,709	557,530	52,897
Train-Content	56,609	436,762	13,767
Tune	200	3,258	1,442
Test	649	10,334 (545)	4,355
ENGLISH			
Train	45,709	723,399	26,747
Train-Content	56,609	403,162	19,791
Tune	200	4377	1657
Test	649	13,484 (231)	3,220

TURKISH	**Morph-emes**	**Unique Morp.**	**Morp./ Word**	**Unique Roots**	**Unique Suff.**
Train	1,005,045	15,081	1.80	14,976	105
Tune	6,240	859	1.92	810	49
Test	18,713	2,297	1.81	2,220	77

and disambiguated form.

Chapter 4

EXPERIMENTS WITH PHRASE-BASED STATISTICAL MACHINE TRANSLATION

To improve the translation quality, we focus on two points: one is obtaining more reliable alignments and the other is the post-processing of decoder output. This chapter studies improving word alignments whereas post-processing is explained in Chapter 5.

We perform different experiments with various representations of Turkish and English texts to exploit the morphology on both sides. We augment data with the dictionaries and the content words to bias the content word alignments and with phrase tables to improve the phrase alignments. We try some derivational morphology segmentation and local reordering on English side.

4.1 Word (Baseline) Representation

As a baseline system, we used morphemic representation of English and Turkish sentences as "full" words. An example is;

T: `kat+hl+ma ortaklık+sh+nhn uygula+hn+ma+sh , ortaklık anlaşma+sh çerçeve+sh+nda izle+hn+yacak+dhr`
E: `the implementation of the accession partnership will be monitor+vvn in the framework of the association agreement`

4.2 Morphemic Representation

We experimented with different morphologically segmented and disambiguated versions of parallel texts to maximize the alignment and consequently translation quality. The use of morphemic representation is particularly important in order to uncover relations between Turkish morphemes and function words on one side and English morphemes and function words on the other side, in addition to relations between open class content words. As morphemes are separated from the root words and allomorphs are abstracted to their lexical forms, the statistics combine and the data sparseness problem is less acute.

We trained the same system with four different morphemic representations of the parallel texts. The decoder now produced the translations as a sequence of root words and morphemes. The surface words were obtained by just concatenating all the morphemes following a root word (until the next root word) taking into account just morphographemic rules but not any morphotactic constraints. As expected this *morpheme-salad* produces a *word-salad*, as sometimes wrong morphemes are associated with incompatible root words violating many morphotactic constraints. This output needs a further post-processing to repair such words.

4.2.1 Full Morphological Segmentation

In the full morphological segmentation, root words and bound morphemes/tags of English and Turkish sentences are represented as separate tokens. Above example is represented as follows;

T: `kat +hl +ma ortaklık +sh +nhn uygula +hn +ma +sh ,ortaklık anlaşma +sh çerçeve +sh +nda izle +hn +yacak +dhr`

E: `the implementation of the accession partnership will be monitor +vvn in the framework of the association agreement`

4.2.2 Root + Morphemes Representation

Certain sequence of morphemes in Turkish texts are translations of some continuous sequence of functional words and tags in English texts and some morphemes should be aligned differently depending on the other morphemes in their context. Therefore, we attempted a selective segmentation of morpheme groups. For example the morpheme `+DHr` in the morpheme sequence `+DHr+mA`, marks infinitive form of a causative verb which in Turkish inflects like a noun; in the lexical morpheme sequence `+yAcAk+DHr` usually maps to `it/he/she will`.

The aim of this process was two-fold: it lets frequent morpheme groups behave as a single token and help training word alignments with identification of some of the phrases. Also since the number of tokens on both sides were reduced, this enabled GIZA++ to produce somewhat better alignments.

We introduce a representation by just separating the root words from the full form. We emphasize the placement of the root words and test whether the Turkish morpheme groups can map to English morphemes and functional words as a whole. Turkish sentences are represented with roots and combined morphemes. For English

sentences, we used the same representation in full morphological separation. For example the above sentences are;

T: `kat +hl+ma ortaklık +sh+nhn uygula +hn+ma+sh ,ortaklık anlaşma +sh çerçeve +sh+nda izle +hn+yacak+dhr`
E: `the implementation of the accession partnership will be monitor +vvn in the framework of the association agreement`

4.2.3 Selective Morphological Segmentation

A systematic analysis of the alignment files, as shown below, produced by GIZA++ for training sentences showed that certain morphemes on the Turkish side were almost consistently never aligned with anything on the English side; For example, the compound noun marker morpheme in Turkish (**+sh**) does not have a corresponding unit on the English side, as English noun-noun compounds do not carry any overt markers. Such markers were never aligned to anything or were aligned almost randomly to tokens on the English side.

English to Turkish Alignment:
`complete territorial reform and develop the concept of regional and municipal management .`
`NULL ( ) toprak ( 2 ) reform ( 3 ) +sh ( ) +nhn ( ) tamamla ( 1 ) +hn ( ) +ma ( ) +sh ( ) ve ( 4 ) bölge ( 9 ) ve ( 10 ) belediye ( 11 ) yönetim ( 12 )`
`+sh ( ) kavram ( 7 ) +lar ( ) +sh ( 8 ) +nhn ( 6 ) geliş ( 5 ) +dhr ( ) +hl ( ) +ma ( ) +sh ( ) . ( 13 )`
Turkish to English Alignment:
`toprak reform +sh +nhn tamamla +hn +ma +sh ve bölge ve belediye yönetim +sh kavram +lar +sh +nhn geliş +dhr +hl +ma +sh .`
`NULL ( 3 8 14 23 ) complete ( 4 5 6 7 18 22 ) territorial ( 1`

) reform (2) and (9) develop (19 20 21) the ( ) concept (15) of (16 17) regional (10) and (11) municipal (12) management (13) . (24)

Since we perform derivational morphological analysis on the Turkish side but not on the English side, we also noted that most verbal nominalizations on the English side were just aligned to the verb roots on the Turkish side and the additional markers on the Turkish side indicating the nominalization and various agreement markers etc., were mostly unaligned.

We listed the Turkish features and their unalignment percentages from the full morphological segmented corpus. In this analysis we preferred features instead of morphemes, as some features are represented exactly the same in the morpheme level such as both `+Cop(Copular)` and `+Caus (Causative)` represented with `+dhr` and `+Inf2 (Infinite)` and `+Neg (Negative)` represented with `+ma`. Table 4.1 shows some highly frequent morphemes and their unalignment percentages.

We selected unaligned morphemes with unalignment percentage over %80 and attached such morphemes (and in the case of verbs, the intervening morphemes) to the root. Otherwise, we kept other morphemes, especially any case morphemes, still separate, as they almost often align with prepositions on the English side quite accurately. It should be noted that what to selectively attach to the root should be considered on a per-language basis; if Turkish were to be aligned with a language with similar morphological markers, this perhaps would not have been needed. Again one perhaps can use methods similar to those suggested by Talbot and Osborne [57].

In this case, the Turkish word above would be represented by a root and some groups of morphemes. For English sentences, we used the same representation in full morphological separation. For example the above sentences are;

T: kat+hl+ma ortaklık+sh +nhn uygula hn+ma+sh ,ortaklık anlaşma+sh çerçeve+sh +nda izle+hn +yacak +dhr

Feature(Morpheme)	Count	Unalignment Percentage
+p3sg (+sh)	152618	93.41
+a3pl (+lar,+larh)	66837	20.92
+loc (+da,+nda)	45620	58.26
+pass (+hl)	44843	54.95
+gen (+nhn)	42214	85.20
+inf2 (+ma)	41835	86.58
+cop (+dhr)	29664	75.20
+caus (+dhr)	20732	83.78
+prespart (+yan)	17216	77.96
+narr (+mhs)	16294	19.85
+ness (+lhk)	12557	77.05
+ins (+yla)	8832	70.09
+pastpart (+dhk)	7581	54.21
+neg (+ma)	6148	42.20
+fut (+yacak)	5105	32.67
+acquire (+lan)	5101	80.96

Table 4.1: Unalignment probabilities for Turkish morphemes

E: `the implementation of the accession partnership will be monitor +vvn in the framework of the association agreement`

4.3 Augmenting Data with Content Words

In order to overcome the disadvantages of the small size of our parallel data, we experimented with ways of using portions of the training data as additional training data.

We add the open class content word training data both baseline and morphemic representation as a bias for content word matchings. By doing this, we expect EM algorithm to learn content matchings better.

4.4 English Derivational Morphology

When processing our parallel data, we did not attempt to do derivational morphology on the English side as the tagger did not perform any further morphological decomposition other than stemming. Even it is not as complicated as Turkish, English morphology also uses derivational morphemes to form new words similar to Turkish. For example;

(zor-zorluk) (difficult-difficulty)
(köy-köylü) (village-villager)
(değer-değersiz) (worth-worthless)
(arkadaş-arkadaşlık) (friend-friendship)

English employs both prefixes and suffixes to make derivations such as `friend +ship, develop+ment, un+tie, un+happy+ness`. Previous representational experiments are carried out by just exploiting inflectional morphology on the English texts. In this work, we describe some initial experiments with English derivational morphology. In order to gauge if such additional information could provide any enhancement, we used the CELEX database[1] to split derivations of English words into morphemes.

We did two different experiments for English derivational morphology. In the first one, we selected high frequent words and segmented them into their derivational morphemes. In the second approach, we tried to segment English words in a similar fashion to Turkish segmentation. For the morphemes that have one representation in Turkish, we assumed that they are allomorphs of an lexical morpheme and abstracted them into their Turkish counterpart. For example, we grouped the morphemes `+ship`, `+ness` and `+ity` and represented them with the Turkish morpheme `+lhk`. Similarly, we grouped `+ion`, `+ation` and `+ition` and represented then with

[1] http://www.ru.nl/celex/

Turkish nominalization morpheme +ma. In these experiments, Turkish is represented with selectively segmented sentences.

First Approach E: `implement +ation of access +ion partner +ship monitor +vvn framework associate +ion agree +ment`

Second Approach E: `implement +ma of accede +ma partner +lhk will +vb monitor +vvn framework in of associate +ma agree +ma`

4.5 Reordering

Language pairs rarely share a common word order. The differences between word orders complicate getting good word alignments, and drop phrase extraction and target translation quality. To match the target language word order, researchers force SMT decoders to employ reordering schemes as they are generating the target language. However, decoding should be an efficient and fast step, so most decoders use very simple reordering schemes that support monotonic translation and generally tend to penalize candidates with long distance reorderings [15, 34]. Phrase-based models typically have a simple distortion model that reorders phrases independently of their content [15, 16], or not at all [29, 30].

It has been observed that one gets better alignments and hence better translation results when the word orders of the source and target languages are more or less the same. When word orders are systematically different, researchers have tried systematically reordering the tokens of source sentences to an order matching or very close to the target language word order, so that alignments could be very close to a monotonic one. Thus instead of forcing the decoders to employ reordering schemes, the source sentences are similarly reordered and then decoded with the decoder employing a hopefully simpler reordering models.

At the constituent level, although the dominant constituent order in Turkish Subject-Object-Verb, essentially all possible orders are possible without any substantial formal constraints, depending on the discourse context. On the other hand, English is essentially Subject-Verb-Object. Moreover, Turkish and English show more local differences in phrase formations. Turkish verb phrases are formed by means of suffixes attached to a root. In English, it is basically formed by using functional words, personal pronouns and possessive determiners included before verb. In order to make the source and target language word orders closer, one approach is to use the morpho-syntactic information and reorder the source language before word alignment in a preprocessing step. Reordering target language is not preferred as an additional post-processing needed.

In order to make the source and target language word orders closer, one approach is to use the morpho-syntactic information and reorder the source language before word alignment in a preprocessing step.[2] The whole point is to bring the relative orders of tokens to a reasonably monotonic state hoping that it would help with alignment and eventually with decoding. So the trained system expects reordered sentences (which do not have to be valid English sentences) and then produces Turkish sentences which is afterwards compared with BLEU to the references.

Our goal is *not to attempt a full reordering at the sentence constituent level*. Instead, we have a more modest goal of a very local and limited source word reordering for a certain class of phrases. If the word order in an English phrase has a more or less monotonic alignment with the morpheme order of the corresponding morphologically marked Turkish word we hope to obtain more reliable phrase alignments.

To handle the word reordering, we offer a pattern extraction method depending on the part-of-speech tags in the English texts. Our approach learns rewrite patterns from source language texts statistically especially for prepositional and verb phrases.

[2]Reordering target language is not preferred as an additional post-processing needed.

We then apply rewrite patterns on the training and test data as a preprocessing step. This procedure is language and context specific allow allows us to focus on the relevant transformations. We use the fully tagged and unsegmented English texts for extraction procedure. Following example shows phrase alignments between Turkish and fully tagged English sentences.

E: [[the+dt implementation+nn]$_1$ of+in$_2$ [the+dt accession+nn partnership+nn]$_3$]$_4$ [will+md be+vb monitor+vv+vvn]$_5$ [[in+in the+dt framework+nn]$_6$ of+in [the+dt association+nn agreement+nn]$_7$]$_8$.
T: [[kat+hl+ma ortaklık+sh]$_3$ +nhn$_2$ [uygula+hn+ma+sh]$_1$]$_4$, [[ortaklık anlaşma+sh]$_7$ [çerçeve+sh+nda]$_6$]$_8$ [izle+hn+yacak+dhr]$_5$.

For the Turkish English pair, the types of possible transformations are rather limited: The PP[3] NP[4] reordering and verb complex ordering are the two major types: since verbs and nouns are the only productively inflecting/deriving word classes. So some linguistic rule based approach is probably quite suitable. To motivate such reordering we present the following examples:

- Turkish noun forms with cases other than nominative case (which is the default case when no case suffixes are present) typically correspond to (parts of) prepositional phrases in English. For example, in

  in$_1$ my$_2$ long$_3$ story$_4$+s$_5$ $\leftrightarrow$ uzun$_3$ hikaye$_4$+ler$_5$+im$_2$+de$_1$

  a reordering of the function words in the English prepositional phrases leads to

  long$_3$ story$_4$+s$_5$ my$_2$ in$_1$ $\leftrightarrow$ uzun$_3$ hikaye$_4$+ler$_5$+im$_2$+de$_1$

[3]PP denotes prepositional phrases
[4]NP denotes noun phrases

in which both the source (word) and the target (morpheme) tokens are monotonically aligned.

- English auxiliary verb complexes and infinitive forms can be reordered to monotonically align to Turkish verb forms or Turkish infinitives. For example, in

  $$\texttt{will}_1\ \texttt{be}_2\ \texttt{monitor}_3\texttt{+ed}_4 \quad \leftrightarrow \quad \texttt{izle}_3\texttt{+n}_4\texttt{+ecek}_1\texttt{+tir}_2$$

  a reordering of the auxiliary verb components leads to

  $$\texttt{monitor}_3\texttt{+ed}_4\ \texttt{will}_1\ \texttt{be}_2 \quad \leftrightarrow \quad \texttt{izle}_3\texttt{+n}_4\texttt{+ecek}_1\texttt{+tir}_2$$

  in which again both the source and the target tokens are monotonically aligned.

4.5.1 Related Work

A number of previous studies have addressed the use of morpho-syntactic information in reordering schemes. Brown et al. [11] reorder phrases by the help of an analysis preprocessor. Xia and McCord [58] derive reordering patterns from word alignments and use these patterns in monotonic decoding. Niessen and Ney [59] focus on reordering separated German verb prefixes and question inversion by using POS tags. Collins et al. [60] uses hand-written rules for reordering German clauses. Popovic and Ney [61] reorder adjectives in English-Spanish SMT by using POS tags. Recently, Wang et al. [62] showed improvement on - a language pair with very different word orderings- Chinese-English SMT by using Penn Chinese Treebank phrase types and Zwarts and Dras [63] reorder source sentence words by minimizing the dependency distance between the head and dependent.

4.5.2 Prepositional Phrases

English prepositional phrases consist of prepositions, pronouns and/or possessives preceding the root words. In Turkish on the other hand, morphemes compounding to English prepositions are attached to the end of the root word. For example;

in `the framework (çerçeve +sh` +**nda**)
from `20 June 1995 (20 Haziran 1995` +**ndan**)
to `Turkey ( Türkiye` +**dan**)
`15 %` **of** `the meda bilateral appropriation (meda iki+lh yardım+lar +sh`+**nhn** `% 15 +yh)`

To investigate the impact of the local reordering, we selected nine prepositions (`of, in, from, to, for, on, at, under, into`) occurring with high frequency on the English side of the training data.[5]

We are not actually parsing the sentences in the sense of a full parsing. Our sentences are already tagged with parts of speech and we are essentially bracketing short PP's of up to 4 tokens on the English side only using part-of-speech information. The idea here is that a a PP with one determiner/possessor and possibly a plural marker would most of the time have the same components of a case-marked Turkish noun with a possessor and a plural marker: e.g., `in my drawer +NNS` ↔ `çekmece+ler+im+de`.

We extract rewrite patterns as follows:

- For each selected phrase type, we search source language texts for the rewrite patterns. This step differs for each phrase type and is explained in detail below;

[5]All our tests with `with` failed to improve the results (one possible reason may be that the English "with" does not always correspond to case markers in Turkish but may also correspond to conjunctions and present participles and full case-marked nouns)

- We count the occurrences of patterns and remove patterns having low frequencies, including punctuations and linguistically meaningless patterns. For example, `+dt` *determiner* `+cc` *conjunction*;
- We start from the longest pattern, process the source language text in a left to right fashion and reorder phrases that match the patterns.

For the prepositional phrases, except the preposition `of`, we search patterns in the form of $PP = preposition\ tag_1 tag_2 \dots tag_i$ up to length 4. For nouns, the root and any plural marker is kept, any preceding possessive pronouns is placed after these two and the preceding preposition is placed after the possessive pronoun.

The case of `of` presents special difficulties: `of` maps to an explicit case morpheme not as frequently as the others, for example, in NPs like `The Queen of England` the `of` do not map to a genitive morpheme on the Turkish equivalent of England. Moreover, noun phrases on both sides of `of` have to be identified and swapped, that is $\texttt{NP}_1$ `of` $\texttt{NP}_2$ is reordered to $\texttt{NP}_2$ `of` $\texttt{NP}_1$, to match the ordering on the Turkish side. Note that if the first $\texttt{NP}_1$ is part of a prepositional phrase, it has to be reordered first. The situation becomes more complicated with any errors in the bracketing of the two NPs on each side.

For preposition *of* we search patterns in the form of $of\_PP = tag_1 tag_2 \dots tag_i\ of\ tag_1 tag_2 \dots tag_j$ up to length 4. For preposition `of`, the first step of extraction procedure should obtain patterns also by checking preceding tags. We then swap the preceding and succeeding tag groups.

Table 4.2 shows some top rewrite patterns.

The reordered English sentence is as follows;

E: the+dt accession+nn partnership+nn of+in the+dt implementation+nn will+md be+vb monitor+vv+vvn the+dt association+nn agreement+nn of+in the+dt framework+nn in+in

Before → After	**Frequency**
+in +dt +nn → +dt +nn +in	2909
+in +dt +jj +nn → +dt +jj +nn +in	1708
+in +nn → +nn +in	1465
+in +dt +nn +nn → +dt +nn +nn +in	584
+in +cd → +cd +in	491
+from +cd +nn → +cd +nn +from	71
+from +dt +np +np → +dt +np +np +from	94
+from +dt +nn+nns → +dt +nn+nns +from	53

Table 4.2: Rewrite Patterns for Some Prepositional Phrases

4.5.3 Verb Phrases

English verb phrases may contain preceding negation and auxiliary verbs, a main verb and succeeding tense suffixes. Similar to the prepositional phrases, Turkish verbs are formed by attaching tense/negation/auxiliary morphemes to the end of the root word. For verb phrases, we search the texts to find the patterns in the form of $VP = tag_1 tag_2 \ldots tag_i \; root\_word + vv + tense\_suffix \; i = 1..4$. In the preprocessing step, main verbs optionally containing tense suffix is moved to the beginning of the phrase. The following example shows the English sentence after verb phrase reordering.

E: `the+dt accession+nn partnership+nn of+in the+dt implementation+nn monitor+vv+vvn will+md be+vb the+dt association+nn agreement+nn of+in the+dt framework+nn in+in .`

4.5.4 The Determiner `the`

In addition to these local reorderings, we remove the determiner `the` from the English side as there is almost never a counterpart on the Turkish side except a few cases. On the contrary, the determiner `a` always has a counterpart.

As a result of these local reordering and removal of *the*, the aligned sentence pair given earlier (and with selective segmentation already applied), along with aligned tokens coindexed, looks like

E: `accession+nn partnership+nn of+in implementation+nn monitor+vv+vvn will+md be association agreement of framework in .`

Note that the top level phrasal constituent orders are still different (Subject-Object-Verb vs Subject-Verb-Object) but within each constituent, the alignments are monotonic, to the extent possible.

After the reordering process a sample full morphological segmented sentence pair with word alignment indixes is as follows:

T: $[[\texttt{kat+hl+ma ortaklık+sh}]_3$ $\texttt{+nhn}_2$ $[\texttt{uygula+hn+ma+sh}]_1]_4$, $[[\texttt{ortaklık anlaşma+sh}]_7$ $[\texttt{çerçeve+sh+nda}]_6]_8$ $[\texttt{izle+hn+yacak+dhr}]_5$.

E: $[[\texttt{accession+nn partnership+nn}]_3$ $\texttt{of+in}_2$ $[\texttt{implementation+nn}]_1]_4$ $[$ $\texttt{monitor+vv+vvn will+md be+vb}]_5$ $[[\texttt{association+nn agreement+nn}]_7$ $\texttt{of+in}$ $[\texttt{framework+nn in+in}]_6]_8$.

4.6 Experiments

4.6.1 Experimental Setup

We employ the phrase-based statistical machine translation framework [15], and use the Moses toolkit [34]with GIZA++ tool [64] for word-based alignments[6] and the SRILM language modelling toolkit [65], and evaluate our decoded translations with the BLEU metric [7], using a single reference translation.

As the average Turkish word in running text has between 2 and 3 morphemes we limited ourselves to 40 words in the parallel texts in order not to exceed the maximum number of words recommended for GIZA++ training.

The test set was selected from the complete data uniformly by extracting every 100th sentence until we had 650 sentences.[7] We also use multiple similar test sets in the first experiments and found that they varied by about 1 point in results and did not pursue multiple test sets after that.

In all experiments, the representation of the Turkish train and reference sentences were the same. The test sets were also modified accordingly on the English train sentences whenever applicable and the Turkish candidate translation was generated with the appropriate representation. For example, for the selective morphological segmentation, all sentences in the test and the train on the Turkish side were selectively segmented; and respectively for the other representations.

For the language model, we used the complete Turkish sentences from the training data with an additional monolingual Turkish text of about 100K sentences coming from news texts which we can consider as out of domain with respect to the training parallel texts.[8] A 5-gram morpheme-based language model was constructed for Turkish (to be used by the decoder). The decoder also produced 1000-best can-

[6]The phrase table was extracted using a maximum phrase size of 7.
[7]We dropped one sentence as its length is too long.
[8]Test data is excluded, not to bias the decoding.

Experiment/Decoder Parameters	**BLEU**	**BLEU** *w Content Words*
Word-based Baseline/Default Parms	NA	16.13
Word-based Baseline/Modified Parms	20.16	19.77

Table 4.3: BLEU Results for the baseline representation

didate translations and then via a script combined the translation model score and the language model score through a small set of weight combinations to see where we would hit the maximum BLEU. We used the best combination with 0.4 for translation model score weight and 0.6 for language model score weight to evaluate the test set with and rescored using weighted combinationof the 4-gram *word-based* language model score and the translation score produced by the decoder.[9]

For the BLEU evaluation, all representations were converted to the word-based representation by concatenating the morphemes to the previous root group.

4.6.2 Results

In the first set of experiments we focus on the representation of Turkish sentences. Tables 4.3 and for 4.4 show experimental results for baseline and morphemic representations, respectively. The test corpus was decoded with two different parameters, with default parameters and modified parameters `(-dl -1)` to allow for long distance movement and `(-weight-d 0.1)` to avoid penalizing long distance movement. We arrived at this combination by experimenting with the decoder to avoid the almost monotonic translation we were getting with the default parameters. These parameters boosted the BLEU scores substantially compared to default parameters used by the decoder.

The decoded output and evaluation results indicate that the standard word-based models for English to Turkish statistical machine translation are quite far

[9]The combination weights were optimized on the tune corpus.

Experiment/Decoder Parameters	**BLEU**	**BLEU w Content Words**
Full morphological Segmentation/Default Parms	13.55	NA
Full morphological Segmentation/Modified Parms	20.22	21.47
Full morphological Segmentation/Modified Parms + Rescoring	21.01	22.18
Root+Morphemes Segmentation/Modified Parms	NA	20.12
Selective Morphological Segmentation/Modified Parms	NA	23.47
Selective Morphological Segmentation/Modified Parms +Rescoring	NA	**24.61**

Table 4.4: BLEU Results for the morphemic representations

from accurate translations into Turkish even with modified parameters. Moreover, augmenting with content words lowers performance. This result is not that interesting; words are represented in the baseline form therefore content words treated as new words for the training data, and cannot help the statistics of word forms.

In the morphemic representation, we observed that the default decoding parameters used by the Moses decoder produces much worse results especially for the fully segmented model. Although some of this may be due to the (relatively) small amount of parallel texts we used, it may also be the case that splitting the sentences into morphemes can play havoc with the alignment process by significantly increasing the number of tokens per sentence especially when such tokens align to tokens on the other side that is quite far away.

Once we recognized that the default parameters were giving very inferior results, we opted not to pursue the default parameters for other representations. The use of the content words as a bootstrapping dictionary significantly increases BLEU scores more than 1 points by constraining possible root word alignments, or boosting correct alignments.

We can conclude that morphemic representation can locate more root words and better word orders correctly than baseline model. The best BLEU results

Experiment	BLEU
frequency threshold > 50	23.71
frequency threshold > 1000	24.11

Table 4.5: BLEU Results for English derivational Morphology with frequencies

Morpheme Abstraction	BLEU
1. -ship,-ness,-ity → +lhk	24.70
2. -ship,-ness,-ity → +lhk +Turkish full segmented	18.90
3. 1 + -ion,-ation, -ition → +ma	24.12
4. 1 + -ion,-ation, -ition → +ma +Turkish full segmented	18.43
5. 2 + in-,-less → +shz, -ous → +lh, -en → +mak	24.00
6. 5 + -al, -ial → +sal, -ive, -ative → +ch, -able, -ible → +yabil	23.43

Table 4.6: BLEU Results for English derivational Morphology with abstraction

are obtained with selective morphological segmentation as 24.61 and represents a relative improvement of 23%, compared to the respective baseline of 19.77.

Our further experiments were only executed on top of the results of the best performing representation (selective morphological segmentation) and modified parameters. The training corpus was augmented with the content word parallel data in all of the following experiments.

For the first set of English derivational morphology experiments, we selected two thresholds, 50 and 1000 to see the effect of frequency of words on the English derivational segmentation. Table 4.5 shows that both of the experiments produced lower scores than the previous top scoring system.

Secondly, we selected 4 different groups of morphemes and abstracted them into Turkish morphemes to collect the statistics. Morpheme groups are abstracted as: `(-ship,-ness,-ity)` to `+lhk`, `(-ion, -ation, -ition)` to `+ma`, `(in-, -less)` map to `+shz`, and `(-ous)` to `+lh`. Table 4.6 shows the BLEU scores for the above experiments with both full and selectively segmented Turkish sentences.

Decomposing words into morphemes similar to Turkish full segmentation lowers the BLEU scores for two reasons: first, English does not have a systematic and regular derivational process. For example, the morpheme `-en` that derives verb from adjective such as `(weak-weaken)`, `(short-shorten)` cannot be applied all adjectives such as `long (lengthen)`, `big (grow)` and `thin (thin)`. Second, some derivational processes do not have a counterpart in the Turkish sentences. For example, the morpheme `-er` that derives noun from verb such as `(kill-killer)`, `(teach-teacher)` do not match any morpheme in the Turkish part as Turkish translations of these words are `katil`[10] and `öğretmen`.[11] Therefore, the English sentences and Turkish sentences cannot be parallelly segmented. As we did not observe any improvements in the BLEU score compared to our previous best results, we did not use English derivational morphology in the subsequent examples.

For reordering experimentation, we considered different subsets of the transformations as seen below:[12]

- in *prep1*, prepositional phrases headed by all prepositions except *of*, were reordered.
- in *prep2*, prepositional phrases headed by all nine prepositions were reordered.
- in *inf*, infinitive verb constructs (headed by `to`) were reordered
- in *the*, the determiner `the` was dropped
- in *verb*, all auxiliary verb sequences were reordered

Table 4.7 shows the results of experiments on the top scoring system (24.61) with various combinations of the transformations above. The best results have been

[10]Although the word `killer` can be translated as `öldürücü(öl+dhr+yhch)` or `öldüren (öl+dhr+yan)` `katil` is the most common translation.

[11]This word has two analysis as `öğret+ma+hn` and `öğretmen`. Morphological disambiguator always selects the second analysis.

[12]The local transformations were restricted to sequences occurring more than 10 times, with length up to 4 tokens and did not involve full NP bracketing.

Transformation	BLEU
of+the	23.58
prep2+the+inf	24.12
verb+of+the	24.05
verb+prep1+inf+the	24.55
verb	24.56
verb+the	24.87
prep1+the+inf	**25.35**

Table 4.7: BLEU Results for various reordering schemes

obtained with the local ordering of the prepositional phrases headed by prepositions in the set `prep1`, the removal of the determiner *the* and reordering of the infinitive constructs.

4.7 Some Examples

Below, we present translations of some sentences from the test data along with the literal English paraphrases of the translated and the reference sentences. We also provide the decoder input and some remarks about the translation produced.

Sentence 1:

Input: `promote protection of children's rights in line with eu and international standards`

Decoder Input: `promote protection of child +nns +pos right +nns line in with eu and international standard +nns`

Translation: `çocuk hak+lar+sh+nhn korunma+sh+na yönelik ab ve uluslararası standart+lar+yla uyum+lh+dhr`

Word Representation: `çocuk haklarının korunmasına yönelik ab ve uluslararası standartlarla uyumludur`

Literally: `it is compatible with the eu and international standards regarding the protection of children's rights`

Reference: ab ve uluslararası standartlar doğrultusunda çocuk haklarının korunmasının teşvik edilmesi
Literally: in line with eu and international standards , promote protection of children's rights

Although the translation does not completely match the reference, it displays one interesting aspect: the underlined morpheme is actually case marker that functions as the subcategorization marker for the following root word with which it forms a translation phrase. This case marker gets attached to the end of the previous word, so that resulting case-marked noun is well-formed and satisfies the subcategorization constraint. This is an example of the decoder creating a morphologically legitimate word by getting the root from one phrase and the morpheme from another phrase.

Sentence 2:
Input: as a key feature of such a strategy, an accession partnership will be drawn up on the basis of previous european council conclusions.
Decoder Input: as a key feature of such a strategy , an accession partnership draw +vvn will +vb up on basis of previous european council conclusion +nns .
Translation: böyle bir strateji+nhn kilit unsur+sh , bir önce+ki avrupa konsey+sh sonuç+lar+sh temel+sh+nda bir katılım ortaklık+sh belge+sh hazırla+hn+yacak+dhr .
Word Representation: böyle bir stratejinin kilit unsuru , bir önceki avrupa konseyi sonuçları temelinda bir katılım ortaklığı belgesi hazırlanayacaktır .
Literally: The lock feature of such a strategy, an accession partnership document based on the previous european council results will be prepared.
Reference: bu stratejinin kilit unsuru olarak , daha önceki ab zirve sonuçlarına dayanılarak bir katılım ortaklığı oluşturulacaktır .

Literally: as a lock feature of this strategy an accession partnership based on earlier eu summit resolutions will be formed .

It clearly has missing, incorrect but also interesting parts: we see that the English key is translated to the Turkish kilit (lock) which is the correct collocational translation. The segment key feature is properly translated as kilit unsur+sh (lock feature), which is the contextually correct idiomatic translation.

Sentence 3:

Input: no development can be reported in the area of free movement of workers .

Translation: işçi+lar+sh+nhn serbest dolaşım+sh alan+sh+nda bir gelişme kaydet+hl+ma+mhs+dhr .

Word Representation: işçilerinin serbest dolaşımı alanında bir gelişme kaydedilmemiştir .

Literally: No development has been recorded in the area of free movement of workers .

Reference: işçilerin serbest dolaşımı alanında rapor edilecek herhangi bir gelişme de yoktur .

Literally: There is no development to be reported in the area of free movement of workers.

The translation is very much acceptable although there does not seem to be much overlap with the reference.

Sentence 4:

Input: the process to meet these priorities has started and mixed progress can be reported

Translation: bu öncelik+lar+hn karşıla+hn+ma+sh süreç+sh başla+mhs ve karış ilerleme kaydet+hl+mhs+dhr

Word Representation: bu önceliklerin karşılanması süreci başlamış ve karış ilerleme kaydedilmiştir

Literally: the process to meet these priorities has started and mix progress has been reported

Reference: bu önceliklerin yerine getirilmesine ilişkin süreç başlamış ve bazı gelişmeler kaydedilmiştir

Literally: the process to fulfill these priorities has started and some improvements has been reported

The decoder output is a very good translation and it has more accurate phrase translations than the reference sentence. The only problematic part is the translation of the word `mixed` (`mix +vvd` in the decoder input) as *karış*).

Sentence 5:

Input: turkey has recently established a working group within the ministry of the interior composed of representatives from several ministries and law enforcement agencies .

Translation: türkiye+da kur+hl+mhs bir çalışma grup+sh+nda içişler+sh bakanlık+sh+nhn temsilci+lar+sh+ndan oluş+yan çeşitli bakanlık+lar ve kolluk kuvvet+lar+dhr .

Word Representation: türkiye'de kurulmuş bir çalışma grubunda içişleri bakanlığının temsilcilerinden oluşan çeşitli bakanlıklar ve kolluk kuvvetlerdir .

Literally: **Reference**: türkiye yakın zamanda , içişleri bakanlığı bünyesinde , çeşitli bakanlıkların ve kolluk hizmeti ifa eden kurumların temsilcilerinden oluşan bir çalışma grubu oluşturmuştur .

Literally: turkey has recently formed a working group within the ministry of the interior composed of representatives from several ministries and law

`enforcement agencies .`

The translation has some short phrase segments with right root words but many morphemes are not attached correctly. The overall sentence can be called as a phrase-salad.

Sentence 6:

Input: `1 . everyone's right to life shall be protected by law .`

Translation: `1 . herkesin yaşama hak+sh kanun+yla koru+hn+hr .`

Word Representation: `1 . herkesin yaşama hakkı kanunla korunur .`

Literally: `1. everyone's living right is protected with law .`

Reference: `1 . herkesin yaşam hakkı yasanın koruması altındadır .`

Literally: `1. everyone's life right is under the protection of the law .`

This example is very interesting from many aspects. Decoder output seems to be a better translation than the reference sentence. Translation correctly attached the morpheme to the word `yaşam` to form the phrase `right to life`. Similarly, phrase `by law` is exactly translated as `kanun+yla` where `kanun` is a synonym of reference word `yasa`.

Chapter 5

POST-PROCESSING OF DECODER OUTPUTS

5.1 Augmenting Data

In order to overcome the disadvantages of the small size of our parallel data, we augment training data with highly reliable phrase table entries that is generated by the training process. The phrase extraction process performs English-Turkish and Turkish-English word alignments using the GIZA++ tool and then combines these alignments with some additional post-processing and extracts "phrases" that are sequences of source and target tokens that align to tokens in the other sequence. Such phrases do not necessarily correspond to linguistic phrases.

Phrase table entries contain the English (e) and Turkish (t) parts of a pair of aligned phrases. Below a portion of the phrase table is shown.

```
enterprise sector ||| işletme sektörü , ||| 0.5 0.08 0.16 0.03
enterprise sector ||| işletme sektörü ||| 0.66 0.08 0.33 0.03
enterprise sector ||| özel sektörün ||| 0.05 0.01 0.16 0.01
enterprise sector ||| özel sektörünün ||| 0.33 0.01 0.16 0.01
```

Iteration-Data Size	*BLEU*	*N-gram precision*
1- 320K	25.56	52.8/29.5/19.8/14.0
2- 593K	26.47	53.7/30.6/20.8/14.9
3- 894K	26.58	53.7/30.5/20.8/14.9
4- 1213K	27.02	54.0/31.1/21.2/15.2
5- 1545K	27.17	54.4/31.3/21.4/15.3
6- 1887K	**27.20**	54.5/31.5/21.3/15.2

Table 5.1: BLEU Results for the phrase table augmentation

```
international passenger ||| uluslararası yolcu ||| 0.83 0.53 0.71 0.55
international registration of ||| uluslararası tesciline ||| 0.33 0.66 1 0.03
international registration ||| uluslararası tesciline ||| 0.66 0.66 1 0.03
international regulation ||| uluslararası düzenlemelere ||| 0.5 0.01 1 0.01
international regulations ||| uluslararası düzenlemelere ||| 0.5 0.10 1 0.023
```

In each line, the first number is $p(e \mid t)$, the conditional probability that the English phrase is e given that the Turkish phrase is t and the third number is $p(t \mid e)$ which captures the probability of the symmetric situation. Among these phrase table entries, those with $p(e \mid t) \approx p(t \mid e)$ and $p(t \mid e) + p(e \mid t)$ larger than some threshold can be considered as reliable mutual translations in that they mostly translate to each other and not much to others. So we extracted those phrases with $0.9 \leq p(e \mid t)/p(t \mid e) \leq 1.1$ and $p(t \mid e) + p(e \mid t) \geq 1.5$ and added them to further bias the alignment process.

On the top scoring system (25.35), we augmented training data iteratively with extracted phrase pairs. Table 5.1 shows the BLEU scores after this augmentation. The BLEU score result after six iterations of this augmentation scheme is 27.20, resulting in a 37.5% relative improvement over the 19.77 baseline, and 7.3% relative improvement over the best previous result after local reordering.

5.2 Word Repair

Generally, the translation output is not error-free and contains many morphological and/or syntactic errors such as terminology errors, preposition errors, modifiers and word form errors and/or word order errors. Beside these errors, because of the morphemic representation we face many morphotactic and morphographemic errors. The main problem of the morphemic representation is the placement of morphemes. In the translation output, root words can be determined correctly, however the morpheme sequence following the root can have errors. While decoding, some morphemes are deleted and/or some spurious morphemes are attached to the root words which needs morphological correction. Morpheme ordering/translation is a very local process and the correct sequence should be determined locally although the existence of morphemes could be postulated from sentence level features during the translation process. Despite the decoder can generate reasonable sequence of morphemes, insisting on generating the exact sequence of morphemes could be an overkill. A morphological generator could take a *root word* and a *bag of morphemes* and generate possible legitimate surface words by taking into account morphotactic constraints and morphographemic constraints, possibly (and ambiguously) filling in any morphemes missing in the translation but actually required by the morphotactic paradigm. Any ambiguities from the morphological generation could then be filtered by a language model.

We attempt to factor out and see if the translations were at all successful in getting the root words in the translations. To analyze this situation, we cleaned up all morphemes and function words from the test and reference sentences scored them. We call the scores as *BLEU-c* not to confuse the results with the word-based representations.

The detailed BLEU results of 27.20, [54.5/31.5/21.3/15.2] for our best performing model, indicates that only 54.5% of the words in the candidate translations are determined correctly. However, when all words in both the candidate and reference

translations are reduced to roots and BLEU is computed again, we get the BLEU-c results of 32.96, [66.7/38.2/25.2/18.5]. This BLEU-c score with 66.7% 1-gram precision implies we are getting 66.7% of the roots correct in the translations, but only 54.5% of the word forms are correct. Thus by concentrating on getting the morpheme sequences right, we can samewhat improve our results.We analysed erroneous words in three groups; punctuation, malformed words and numbers. Malformed words can be classified into three groups:

1. Morphologically malformed words: words with the correct root word but with morphemes that are either categorically incorrect (e.g., case morpheme on a verb), or morphotactically incorrect (e.g., morphemes in the wrong order). Words in this class would be rejected by our morphological generator. In the below example, morphological generator detects the word `genel+da+yan` as UNKNOWN.

   Translation: `genel+da+yan , mamul mal+lar gümrük birlik+sh +nhn iç+sh+da serbest+ca dolaş+makta+dhr .`
   Reference: `genel ol+yarak , sanayi ürün+lar+sh , gümrük birlik+sh çerçeve+sh+nda serbest dolaşım+da bulun+makta+dhr .`

2. Morphologically well-formed words which are out-of-vocabulary relative to the training corpus and the language model corpus. Since Turkish has a very large number of possible word forms, there really are no well-formed words which are out-of-vocabulary, though there may be well-formed words which are extremely low frequency. Words for this case would be accepted by the morphological analyzer but would not be in the vocabulary of the training and language model corpora. We identify these words with the help of a small script. In the above example, word `serbest+ca` is detected as out-of-vocabulary relative to the training and language model corpus.

3. Morphologically well-formed words which are *not* out-of-vocabulary relative to the training corpus and the language model corpus, but do not match the reference. Words `mamul, mal+lar, iç+sh+da, dolaş+makta+dhr` are in this group. we have no way knowing without looking at the reference if a word falls in this class.

5.2.1 Malformed and Out-of-Vocabulary Words

We propose an output correction procedure for malformed and out-of-vocabulary words. Using a finite state model of lexical morpheme structure of possible Turkish words, with morphemes being as the symbols (except for the letters in roots), we use error-tolerant finite state recognition [66] to generate morphologically correct word forms with the same root, but with the morpheme structure up to 2 unit morpheme edit operations *(add, delete, substitute, transpose morphemes)* away.

As an example, for the sentence `seçim yasak+hl ilan+yacak et+hl+dh`, the words `yasak+hl` and `ilan+yacak` are detected as malformed words. For instance, the word form (in lexical morpheme representation) `ilan+yacak` is malformed and possible correction at distance 1 are {`ilan, ilan+sh, ilan+nhn, ilan+nhn+ya`}. We convert the sentence to a lattice representation replacing each malformed with the correct alternatives as shown in Figure 5.1. For simplicity we just show a small subset of possible words and for readability, we use surface forms of the words in the following examples.

The resulting lattice is then rescored with the morpheme and word language models separately to pick the best alternative for sentence as shown below, with log probabilites assigned by the language model.[1]

```
 -12.36 seçim yasağı ilan edildi
-14.4454 seçim yasakları ilan edildi
```

[1]In this step, the morpheme-based language model performed better than the word-based-language model.

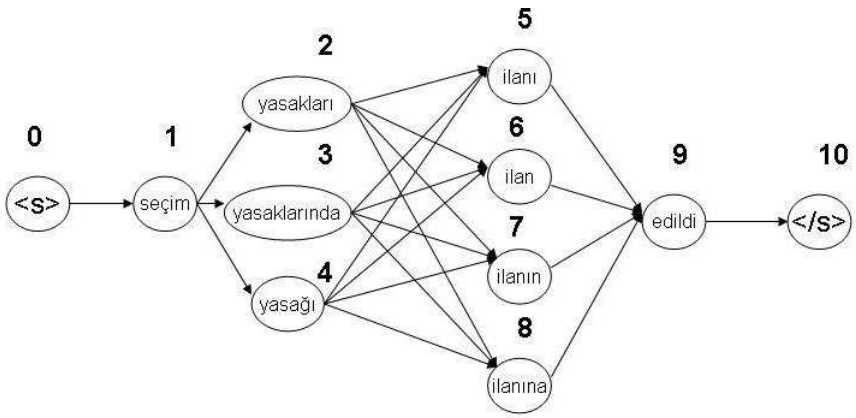


Figure 5.1: A lattice example for correcting malformed words

```
-17.955 seçim yasağı ilanı edildi
-18.3198 seçim yasakları ilanı edildi
-18.3837 seçim yasağı ilanına edildi
-18.7484 seçim yasakları ilanına edildi
-18.8244 seçim yasağı ilanın edildi
-19.1892 seçim yasakları ilanın edildi
-inf seçim yasaklarında ilanın edildi
-inf seçim yasaklarında ilanı edildi
-inf seçim yasaklarında ilan edildi
-inf seçim yasaklarında ilanına edildi
```

After repairing malformed words, we apply same procedure to out-of-vocabulary words.

5.3 Experiments

5.3.1 Setup

We use the word repair script to generate the word lattices. We use the SRILM language modelling toolkit's [65] lattice tool to score the lattices, we take the top sentence produces and evaluate our decoded translations using the BLEU measure [7]. For the language model, we use the same morpheme and word language models that used in previous examples.

Additionally, we removed punctuations that take morphemes (e.g `,+dhr`) before correction and normalized out-of-vocabulary numbers (e.g `2004+ya`) by dropping morphemes after correction.

5.3.2 Results

The detailed analysis of the decoded output with reference translations point out that errors generally are caused by some specific morphemes such as `"+dhr", "+sh", "+nhn", "+ya", "+da", "+yh"`. We restricted the possible morpheme changes (deletion, insertion, replacement) with these six morphemes and scored our lattices with both word and morpheme language models. We obtained best scores with word-based language model and 1-distance morpheme changes.

After decoding, 614 words were detected as malformed and out-of-vocabulary. 385 were selected by lattice rescoring and 70 of the words were exactly repaired and matched with reference. 53 of malformed words were repaired but they became 1 or 2 distance far away from the reference. As 221 of the word roots are not in the reference set, lattice scoring does not improve their matching. Table 5.2 shows the BLEU scores after word repair. All in all, word repair provides an additional relative improvement of 1.5% relative improvement (compared to 27.20 after augmenting data) and the final BLEU score represents a relative improvement of 39.7% over the

Experiment	BLEU
Punctuation Cleaining	27.26
+Malformed and Out-Of-Vocabulary Words	27.54
+Number Normalization	**27.60**

Table 5.2: BLEU Results for the Word Repair

Source Length	Count	BLEU (n-gram precision)
1-5	49	79.16 (91.0/82.3/80.0/87.5)
1-10	130	54.08 (75.0/59.5/50.3/45.1)
1-15	214	45.21 (68.7/51.9/41.9/34.8)
1-20	299	37.47 (64.3/44.2/34.1/26.6)
1-30	450	31.40 (59.2/37.4/26.7/19.8)
1-40	553	29.33 (57.0/34.2/23.8/17.4)
1-50	602	27.94 (55.8/32.7/22.4/16.0)
1-100	649	27.60 (55.4/32.1/21.7/15.5)
5-15	179	44.91 (67.5/51.1/41.8/34.8)
5-20	264	37.25 (63.5/43.7/34.0/26.6)
10-20	189	34.69 (61.3/40.8/31.4/24.5)
20-30	164	26.15 (54.5/31.2/20.8/14.8)

Table 5.3: Detailed BLEU scores for various input sentence length ranges

baseline.

Table 5.3 presents detailed BLEU results for various ranges of input sentence length, for our best performing system. As expected, for short sentence up to 15 words the scores are quite high, given the size of the training data. This has been the observation of other researchers in the field for other language pairs. This improvement is basically due to a number of reasons: as the number of possible ways a short sentence can be cut up into phrases is much more limited, this results in a much smaller search space for both translations and reordering. Thus one may get away with much less pruning during decoding. Also the target language models may be more accurate for short sentences.

5.4 An Alternative Evaluation

While word-to-word comparisons in computing n-gram overlaps are meaningful for some language pairs, the BLEU's all-or-none nature of word comparisons can be particularly harsh for a morphologically complex target language when the translation system generates sequences of morphemes that make up target words. When comparing words, BLEU comparing the words can flag a word as a mismatch even a single morpheme does not match although for example the corresponding target and reference morpheme sequences may contain morphemes with very close morphosemantics and are almost interchangeable. Even if the translation is morphosemantically quite acceptable, the words are assumed not matching. For example word groups;

> `gel+hyor (he is coming)` vs. `gel+makta (he is (in a state of) coming))` are essentially the same. On a scale of 0 to 1, one could rate these at about 0.95 in similarity.

> `gel+yacak (he will come)` vs. `gel+yacak+dhr (he will come)` in a sentence final position. Such pairs could be rated perhaps at 0.90 in similarity.

> `gel+dh (he came)` vs. `gel+mhs (he came)`. These essentially mark past tense but differ in how the speaker relates to the event and could be rated at perhaps 0.70 similarity.

Considering such cases as a complete mismatch downgrades the performance of the system even though it gets most of the morphemes correctly. Because of this way of calculation, the scores assigned by BLEU generally do not reflect the right performance for language like Turkish.

To overcome this problem, there is a need of a weighted calculation of words where stems and morphemes are counted separately. As the occurrence of the right root is more important than the occurrence of right morpheme sequence, the output still can be closer to the right root and wrong/missing/spurious morpheme sequence.

The following candidate and reference translations (with both word-level (W) and lexical-morpheme (L) representations) exemplify the problem more acutely.

> **Candidate:** `iki aile arasındaki husumet ve kavga uzun yıllardır sürüyordu.`
> **Lexical:** `iki aile ara+sh+nda+ki husumet ve kavga uzun yıl+lar+dhr sür+hyor+dh.`
> **Reference:** `iki aile arasında düşmanlık ve çatışma uzun senelerdir sürmekteydi.`
> **Lexical:** `iki aile ara+sh+nda düşmanlık ve çatışma uzun sene+lar+dhr sür+makta+ydh.`
> **Literally:** `The hostility and fight between two families had been lasting for many years.`

In the candidate translation, 4 of the last 6 words not matching the corresponding word in the reference translation. However, `husumet (enmity)` is a synonym of the reference word `düşmanlık` while `kavga (fight)` is a hyponym of `çatışma (confrontation)` in the Turkish WordNET [44]. Also, the roots `yıl (year)` and `sene` are synonyms in the inflected words `yıllardır (for years)` and `senelerdir`.[2] Finally, the verb of the sentence in the candidate and the reference look different, but the difference is due to the use of the two almost synonymous morphemes. For all practical purposes, the candidate translation sentence renders the same meaning as the reference sentence but BLEU is considered as having a significant mismatch.

[2]Note also that the lexical morphemes also surface differently in these words, due to morhophonological processes such as vowel harmony, etc.

In order to alleviate the shortcoming of strict word-based matching used by the standard BLEU measure for languages like Turkish, we scored our top scoring system with an extension tool, BLEU+ [6], that can perform finer-grained lexical comparison taking into synonymous roots (as in METEOR [38]) and almost synonymous *morphemes*. BLEU+ has four interesting extensions;

- Whenever a WordNET ontology is available, it is possible to match root words based on synonymity.[3] Moreover, hypernyms or hyponyms of a root word can be also included into the scoring.
- Similar to synonym root words, BLEU+ can identify some pre-defined synonymous morphemes such as the lexical morphemes `+hyor` and `+makta`.
- BLEU+ can compute scores only considering the roots, that would give an oracle BLEU score which indicates the maximum score that one would get if the morphemes and their order perfectly correct for each word.
- Similar to previous extension, another oracle score is based on identifying words whose roots are similar but the morphological structure of the words are different. If the morpheme sequences of a reference sentence word can be obtained from the decoded output word, by a small number of morpheme insertions, deletions or substitutions, then it may be worthwhile to identify some of these cases and attempt to correct them. This oracle score gives the maximum BLEU score that we can obtain if we can identify and *fix* all words whose roots are similar but the morpheme structures differ by a small number of edit operations (usually 1 or 2).

Table 5.4 shows the results of evaluating our best result with the BLEU+ tool. We see that taking into account candidate root words which are synonyms, hypernyms or hyponyms of reference root words, and synonymous candidate and reference morphemes, a slight improvement in BLEU can be observed. It should be noted

[3] Assuming the candidate and reference translations are available in a morphemic representation.

Matching Scheme	BLEU+ Score (n-gram precisions)
Default BLEU	27.60 (55.4/32.1/21.7/15.5)
Synonyms/Hyponyms/Hypernyms	27.82 (56.0/32.3/21.9/15.6)
Synonymous Morphemes	27.74 (55.7/32.2/21.8/15.6)
Combined	27.97 (56.3/32.5/22.0/15.7)
Root (oracle)	32.96 (66.7/38.2/25.1/18.5)
Morpheme Correction d=1 (oracle)	32.26 (63.0/37.7/25.5/18.5)
Morpheme Correction d=2 (oracle)	32.87 (65.87/38.2/25.8/18.5)

Table 5.4: BLEU+ scores

that evaluation using BLEU+ is not meant to replace the BLEU evaluation, but are used to provide some hints and insights in what kind of errors at the local word structure level are made and how much one can improve the results by focusing on such errors.

5.5 Some Examples

Below we present translations of some sentences from the test data before and after post-processing step.

Sentence 1:

Input: `3 . the joint committee shall adopt its rules of procedure .`

Translation: `3 . ortak komitedir usul kurallarını kabul edecektir .`

Translation After post-processing: `3 . ortak komite usul kurallarını kabul edecektir .`

Literally: `3 . the joint committee will adopt its rules of procedure .`

Reference: `3 . ortak komite kendi uygulama usullerini tesbit edecektir .`

Literally: `3 . the joint committee will determine its procedures of application .`

In this translation, the word komitedir (komite+dhr) detected as out-of-vocabulary word and repaired as komite which matches the reference word exactly.

Sentence 2:

Input: the conclusions of the copenhagen european council recommended that this amount is substantially increased from 2004 .

Translation: kopenhag ab konseyinin sonuçlarını büyük ölçüde artan 2004e bu miktari tavsiye etmiştir .

Translation After post-processing: kopenhag ab konseyinin sonuçlarını büyük ölçüde artan 2004 bu miktar tavsiye etmiştir .

Literally: ??? it recommended largely increasing 2004 this amount the conclusions of the copenhagen eu council .

Reference: kopenhag avrupa zirvesi sonuçlarında bu miktarın 2004 yılından itibaren önemli ölçüde arttırılması tavsiye edilmiştir .

Literally: in the conclusions of the copenhagen europen summit, it is recommended that this amount is significantly increased from 2004 .

In the translation, the word miktar+yh as malformed word. Although the repaired word miktar+sh does not have a match it is now a morphologically correct word with one distance away to the reference. We also applied number normalization to 2004+ya which results a matching in the reference.

Sentence 3:

Input: the indicators being given for the mid years by provisional population projections .

Translation: göstergelerden geçici yıl ortası nüfus projeksiyonları tarafından verilmiştir .+sh

Translation After post-processing: göstergeler geçici yıl ortası nüfus projeksiyonları tarafından verilmiştir .

Literally: the indicators begin given provisional mid year by population

projections .

Reference: göstergeler , geçici nüfus projeksiyonları tarafından ara yıllar için verilmiştir .

Literally: the indicators being given for the mid years by provisional population projections .

Word göstergelerden(gösterge+lar+ya) is repaired and match exactly the reference word. Punctuation morpheme clean to the token .+sh gives us a one more token that match the reference.

Sentence 4:

Input: 16 . the public prosecutor took a statement from the applicant on 28 august 1995 .

Translation: 16 . başsavcıdan 28 ağustos 1995 tarihinde başvuranın ifadesini almıştır .

Translation After post-processing: 16 . başsavcı 28 ağustos 1995 tarihinde başvuranın ifadesini almıştır .

Literally: 16 . chief prosecutor took a statement from the applicant on 28 august 1995 .

Reference: 16 . savcı , 28 ağustos 1995 tarihinde başvuranın ifadesini almıştır .

Literally: 16 . prosecutor took a statement from the applicant on 28 august 1995 .

Last example, is an interesting example; the word public prosecutor (savcı) is wrongly translated as basavcı (chief prosecutor) with ablative morpheme. The word repair process drops this morpheme that exactly matches the reference sentence form but as the root words are not same, this process has no effect on the final evaluation with BLEU.

Chapter 6

CONCLUSIONS

This thesis presented the development and results of an English-to-Turkish phrase-based statistical machine translation system. This language pair was interesting for statistical machine translation as the target language, Turkish, is a morphologically very rich language, with a subject-object-verb constituent order and essentially has infinite vocabulary, while English is relatively poorer in this respect and has a subject-verb-object order.

To get accurate translations, we focused on two points: (i) obtaining more reliable word alignments and (ii) the post-processing of decoder output. Translation into Turkish involves a variety of processes; for example sometimes a *single word in Turkish needs to be synthesized from the translations of two or more (possibly distant) phrases in English.* We have used morphological preprocessing to identify lexical morphemes on both the Turkish and the English words to alleviate the data sparseness problem but more importantly to uncover relationships between the morphemes on the Turkish side with morphemes and function words on the English. We explored various morpheme representations in order to improve the evaluation scores.

Statistical machine translation systems need substantial amounts of aligned texts from which probabilistic translation models can be trained. This was an important

problem for the Turkish and English pair as we do not have many available sources, for such texts. The dearth of available English-Turkish parallel texts suggested that the available data has to be exploited in various ways to make most use of it. Content words from the training data and highly reliable phrase table entries from previous training steps were used as additional sources.

Our explorations into developing a statistical machine translation system from English to Turkish pointed out that using standard models to determine the correct sequence of morphemes within the words is probably not a good idea. Morpheme ordering is a very local process and the correct sequence should be determined locally though the existence of morphemes could be postulated from sentence level features before the translation process. We reordered English phrases in order to get monotonic phrase alignments and so monotonic morpheme alignments and introduced two different levels of language models: a morpheme-based languge model in decoding for accurate ordering of morhemes and word-based language model in reranking for accurate word ordering.

When Turkish sentences are split into morphemes, an important problem was the use of same decoding mechanism and statistical parameters to handle both the very word-local process of morphotactic ordering, and the more global process of sentence constituent ordering. There was not any mechanism to enforce morphotactics other than language model statistics and this sometimes produced word forms with incorrect structure. Detailed analysis of the errors pointed at a few directions such as word-repair, to improve word accuracy. We offered a word-repair procedure for malformed and out-of-vocabulary words.

Figure 6.1 shows the general structure of the English to Turkish statistical machine translation prototype.

Evaluation of Turkish translation seemed to involve processes that are somewhat more complex than standard evaluation metrics: errors in any translated morpheme or its morphotactic position render the synthesized word incorrect, even though the

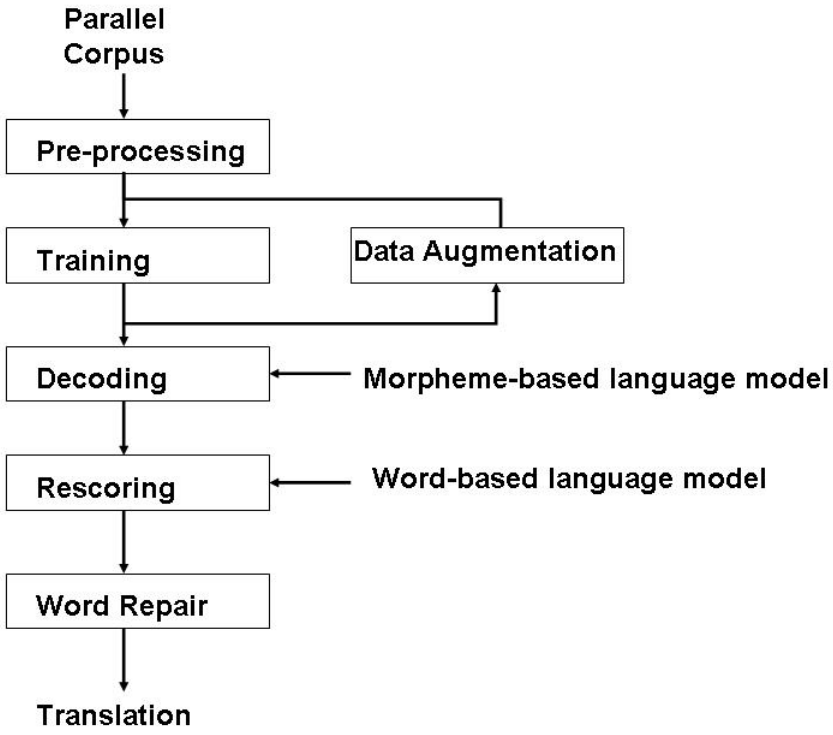


Figure 6.1: English to Turkish statistical machine translation structure

rest of the word can be quite fine. This, though indirectly, implies that BLEU is particularly harsh for Turkish and the morpheme based-approach, because of the all-or-none nature of token comparison when computing the BLEU score. We used BLEU+ tool for a fine-grained evaluation of Turkish sentences.

Major results of our work can be summarized as follows:

1. We have considered various representational schemes to take morphological structure into account. We used the word-based representation as a baseline and presented three different morphologically segmented representations. We decoded test data with these representations with decoder's default parame-

ters which tend to produce monotonic translations and modified parameters to allow a flexible word ordering. All of the morphological representations that decoded with modified parameters performed better than our baseline word representation. Moreover, we have found that employing a language-pair specific morphological representation somewhere between using full word-forms and fully morphologically segmented representations provides the best results. The BLEU score obtained was 23.83 representive 20 % the improvement over the baseline system.

2. We have used Turkish and English WordNets which are aligned via the inter-lingual index as a bootstrapping dictionary to improve root word alignments.

3. We extracted open class content words from the training data to overcome the disadvantages of the small size of parallel texts. We added this content word corpora to the training data. This addition provided some improvement with morphologically segmented representations (by presumably biasing the root word alignments), but not with baseline word-based representation. Using content words as additional data provided a significant boost in BLEU scores and the improvement is 1-1.5 BLEU points on average.

4. We attempted to incorporate English derivational morphology with two different methods. First, we selected high frequent words for segmentation. Second, we selected some English morphemes and abstract them to their Turkish counterparts. However, we did not get any improvement with either of methods.

5. We used morpho-syntactic information for local reordering of certain class of phrases to get a relatively monotonic phrase alignments. We used part-of-speech tags to obtain most frequent and short patterns to reorder English prepositional phrases and infinitive verb structures. In addition, we removed the determiner "the" as it has no translation in the Turkish side. As a result, local reordering gave us approximately 1 BLEU point improvement.

6. We used 5-gram morpheme-based language model for decoding to enforce

	Step	BLEU	% Improvement
0	Word-based Baseline	19.77	
1	Selective Segmented Training Data + Decoding a with Morpheme-based LM	23.83	20.5%
2	(1) + Rescoring with a word-based LM	24.61	24.5%
3	Reordering + (2)	25.35	28.2%
4	Data Augmentation + (3)	27.20	37.6%
5	(4) + Word Repair	27.60	39.6%

Table 6.1: Summary of BLEU Results for all steps of the English-Turkish statistical machine translation

morphotactics constraints and perhaps some very close syntactic constraints. Then, we reranked 1000-best outputs of with with a 4-gram word-based model to enforce longer range constraints. This reranking provided an additional 1 BLEU point.

7. We extracted highly reliable phrase translations from the phrase table and augmented the training data with them that provide additional bias to the alignments and improved the BLEU score about 2 points.

8. On the decoded output, we applied a post-processing procedure that fixes malformed and out-of-vocabulary words. We used lattice-rescoring with word-based language model. After all these steps, we reached a BLEU score 27.60 representive 39.6% improvement over the baseline system.

9. We used the evaluation tool BLEU+ (extension of BLEU metric) that provides various fine-grained analyses of candidate translation by taking into account synonymous roots, and morphemes, and can compute oracle scores to show upper bound performance.

Finally, Table 6.1 shows a summary of the English to Turkish statistical machine translation system steps along with BLEU scores and improvements.

Appendix A

Penn Treebank tags corresponding part of speech

1. CC Coordinating conjunction
2. CD Cardinal number
3. DT Determiner
4. EX Existential there
5. FW Foreign word
6. IN Preposition or subordinating conjunction
7. JJ Adjective
8. JJR Adjective, comparative
9. JJS Adjective, superlative
10. LS List item marker
11. MD Modal
12. NN Noun, singular or mass
13. NNS Noun, plural
14. NNP Proper noun, singular
15. NNPS Proper noun, plural
16. PDT Predeterminer
17. POS Possessive ending
18. PRP Personal pronoun
19. PRP$ Possessive pronoun
20. RB Adverb

21. RBR Adverb, comparative
22. RBS Adverb, superlative
23. RP Particle
24. SYM Symbol
25. TO to
26. UH Interjection
27. VB Verb, base form
28. VBD Verb, past tense
29. VBG Verb, gerund or present participle
30. VBN Verb, past participle
31. VBP Verb, non3rd person singular present
32. VBZ Verb, 3rd person singular present
33. WDT Whdeterminer
34. WP Whpronoun
35. WP$ Possessive whpronoun
36. WRB Whadverb

Bibliography

[1] Weaver, W.: Translation(1949), In Machine Translation of Languages: fourteen essays. MIT Press,Cambridge,MA (1955)

[2] Sagay, Z.: A computer translation from english to turkish. Master's thesis, METU, Department of Computer Engineering (1981)

[3] Hakkani, D.Z., Tür, G., Oflazer, K., Mitamura, T., Nyberg, E.: An english-to-turkish interlingual mt system. In: AMTA. (1998) 83–94

[4] Keyder Turhan, c.: An english to turkish machine translation system using structural mapping. In: Proceedings of the Applied Natural Language Processing, Washington, DC (1997) 320–323

[5] Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT Summit X, Phuket, Thailand (2005)

[6] Tantuğ, A.C., Oflazer, K., El-Kahlout, I.D.: BLEU+: a tool for fine-grained BLEU computation. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco (2008)

[7] Papineni, K., Salim Roukos, T.W., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, Association for Computational Linguistics (2002) 311–318

[8] Hutchins, J.: Machine translation: a brief history. In Koerner, E., Asher, R.E., eds.: Concise history of the language sciences: from the Sumerians to the

cognitivists. Pergamon, Oxford (1995) 431–445

[9] Pierce, J.R., Carroll, J.B., et al.: Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences,National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council, (1966)

[10] Hutchins, J.: Towards a definition of example-based machine translation. In: Proceedings of Workshop on Example-Based Machine Translation, MT Summit X, Phuket, Thailand (2005) 63–70

[11] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Lafferty, J.D., Mercer, R.L.: Analysis, statistical transfer, and synthesis in machine translation. In: Proceeding of TMI: Fourth International Conference on Theoretical and Methodological Issues in MT. (1992) 83–100

[12] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19** (1993) 263–311

[13] Berger, A.L., Brown, P.F., Pietra, S.D., Pietra, V.J.D., Gillett, J.R., Lafferty, J.D., Mercer, R.L., Printz, H., Ures, L.: The candide system for machine translation. In: HLT-NAACL. (1994)

[14] Marcu, D., Wong, W.: A phrase-based, joint probability model for statistical machine translation. In: In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-02), Philadelphia (2002)

[15] Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of HLT/NAACL. (2003)

[16] Och, F.J., Ney, H.: The alignment template approach to statistical machine translation. Computational Linguistics **30** (2004) 417–449

[17] Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, Association for Computational Linguistics (2005) 263–270

[18] Koehn, P., Hoang, H.: Factored translation models. In: EMNLP. (2007)

[19] Brown, P.F., Lai, J.C., Mercer, N.R.L.: Aligning sentences in parallel corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California (1991) 169–176

[20] Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. In: Computational Linguistics. (1991) 177–184

[21] Melamed, I.D.: A geometric approach to mapping bitext correspondence. In: Conference on Empirical Methods in Natural Language Processing. (1996) 1–12

[22] Moore, R.C.: Fast and accurate sentence alignment of bilingual corpora. In: AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users, London, UK, Springer-Verlag (2002) 135–144

[23] Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29** (2003) 19–51

[24] Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Comput. Linguist. **22** (1996) 39–71

[25] Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. (2001) 295–302

[26] Och, F.J.: Minimum error rate training in Statistical Machine Translation. In Hinrichs, E., Roth, D., eds.: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan (2003) 160–167

[27] Vogel, S., Ney, H., Tillmann, C.: Hmm-based word alignment in statistical translation. In: Proceedings of the 16th conference on Computational linguistics. (1996) 836–841

[28] Zens, R., Och, F.J., Ney, H.: Phrase-based statistical machine translation. In Jarke, M., Koehler, J., Lakemeyer, G., eds.: 25th German Conference on Artificial Intelligence (KI2002), volume 2479 of Lecture Notes in Artificial Intelligence (LNAI), Aachen, Germany (2002) 18–22

[29] Zens, R., Ney, H.: Improvements in phrase-based statistical machine translation. In: In Proc. of the Human Language Technology Conf. (HLT-NAACL. (2004) 257–264

[30] Kumar, S., Byrne, W.: A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In: NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. (2003) 63–70

[31] Knight, K.: Decoding complexity in word-replacement translation models. Computational Linguistics **25** (1999) 607–615

[32] Germann, U., Jahr, M., Knight, K., Marcu, D., Yamada, K.: Fast decoding and optimal decoding for machine translation. In: Procedings of ACL-01, Toulouse, France (2001)

[33] Ulrich, G.: Greedy decoding for statistical machine translation in almost linear time. In: Procedings of HLT-NAACL-2003, Edmonton, AB, Canada (2003)

[34] 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., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07) – Companion Volume. (2007)

[35] Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., Sawaf, H.: Accelerated dp based search for statistical translation. In: In European Conf. on Speech Communication and Technology. (1997) 2667–2670

[36] Niessen, S., Och, F.J., Leusch, G.: An evaluation tool for machine translation: Fast evaluation for mt research. In: In Proceedings of the 2nd Internation Conference of Language Resources and Evaluation. (2000) 39–45

[37] Doddington, G.: 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, Morgan Kaufmann Publishers Inc. (2002) 138–145

[38] Banerjee, S., Lavie, A.: Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, Association for Computational Linguistics (2005)

[39] Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM **38** (1995) 39–41

[40] Hankamer, J.: Morphological parsing and the lexicon. In Marslen-Wilson, W., ed.: Lexical Representation and Process. MIT Press (1989)

[41] Barber, C.: English Language. Cambridge University Press (2000)

[42] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Goldsmith, M., Hajic, J., Mercer, R.L., Mohanty, R.: But dictionaries are data too. In: Procedings of the ARPA Human Language Technology Workshop, Princeton, NJ (1993) 202–205

[43] Fellbaum, C., ed.: WordNet, An Electronic Lexical Database. MIT Press (1998)

[44] Bilgin, O., Çetinoğlu, O., Oflazer, K.: Building a Wordnet for Turkish. Romanian Journal of Information Science and Technology **7** (2004) 163–172

[45] Niessen, S., Ney, H.: Statistical machine translation with scarce resources using morpho-syntatic information. Computational Linguistics **30** (2004) 181–204

[46] Yang, M., Kirchhoff, K.: Phrase-based backoff models for machine translation of highly inflected languages. In: Proceedings of EACL. (2006) 41–48

[47] Corston-Oliver, S., Gamon, M.: Normalizing German and English inflectional morphology to improve statistical word alignment. In: Proceedings of AMTA. (2004) 48–57

[48] Lee, Y.S.: Morphological analysis for statistical machine translation. In: Proceedings of HLT-NAACL 2004 - Companion Volume. (2004) 57–60

[49] Zollmann, A., Venugopal, A., Vogel, S.: Bridging the inflection morphology gap for Arabic statistical machine translation. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, New York City, USA (2006) 201–204

[50] Popovic, M., Ney, H.: Towards the use of word stems and suffixes for statistical machine translation. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC). (2004) 1585–1588

[51] Goldwater, S., McClosky, D.: Improving statistical MT through morphological analysis. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada (2005) 676–683

[52] Minkov, E., Toutanova, K., Suzuki, H.: Generating complex morphology for machine translation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, Association for Computational Linguistics (2007) 128–135

[53] Oflazer, K.: Two-level description of Turkish morphology. Literary and Linguistic Computing **9** (1994) 137–148

[54] Yüret, D., Türe, F.: Learning morphological disambiguation rules for Turkish. In: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, New York City, USA (2006) 328–334

[55] Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing. (1994)

[56] Marcus, M., Santorini, B., Marcinkiewitz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics (1993)

[57] Talbot, D., Osborne, M.: Modelling lexical redundancy for machine translation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia (2006) 969–976

[58] Xia, F., McCord, M.: Improving a statistical MT system with automatically learned rewrite patterns. In: In Proceedings of the 20th International Conference on Computational Linguistics (COLING), Geneva, Switzerland (2004) 508–514

[59] Niessen, S., Ney, H.: Statistical machine translation with scarce resources using morpho-syntatic information. Computational Linguistics **30** (2004) 181–204

[60] Collins, M., Koehn, P., Kucerova, I.: Clause restructuring for statistical machine translation. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, Association for Computational Linguistics (2005) 531–540

[61] Popovic, M., Ney, H.: POS-based word reorderings for statistical machine translation. In: 5th International Conference on Language Resources and Evaluation (LREC, Genoa, Italy (2006) 1278–1283

[62] Wang, C., Collins, M., Koehn, P.: Chinese syntactic reordering for statistical

machine translation. In: Proceedings of EMNLP, Prague, Czech Republic, Association for Computational Linguistics (2007) 737–745

[63] Zwarts, S., Dras, M.: Syntax-based word reordering in phase-based statistical machine translation: Why does it work? In: Proceedings of the MT SUMMIT XI, Copenhagen, Denmark (2007) 559–566

[64] Och, F.J., Ney, H.: Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong (2000) 440–447

[65] Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proceedings of the Intl. Conf. on Spoken Language Processing. (2002)

[66] Oflazer, K.: Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. Computational Linguistics **22** (1996) 73–90