SYNTAX-TO-MORPHOLOGY ALIGNMENT AND CONSTITUENT REORDERING IN FACTORED PHRASE-BASED STATISTICAL MACHINE TRANSLATION FROM ENGLISH TO TURKISH

by
Reyyan Yeniterzi
2009

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
August 2009

SYNTAX-TO-MORPHOLOGY ALIGNMENT AND CONSTITUENT REORDERING IN FACTORED PHRASE-BASED STATISTICAL MACHINE TRANSLATION FROM ENGLISH TO TURKISH

APPROVED BY:

to my parents
&
my sister Süveyda

Acknowledgements

I would like to express my deepest gratitude to my advisor, Kemal Oflazer, for his invaluable support, encouragement and supervision. This thesis would not have been possible without his guidance.

I would also like to thank my thesis committee members Dilek Hakkani-Tür, Berrin Yanıkoğlu, Yücel Saygın and Esra Erdem for their valuable comments and suggestions.

I would like to thank Ilknur Durgar El-Kahlout for her help and cooperation throughout the progress of this thesis, and Gülşen Eryiğit for her help with the parser. I am indebted to my fellow colleagues and dear friends Özlem, Ferhan, Burak, Hanife for their endless friendship. I am thankful to Sabancı University faculty and staff for their help and patience throughout these last 7 years.

I would like to thank Erol Çöm for his help during the final submission of this thesis.

The work done in this thesis was partially supported by a seed grant to my advisor by the Qatar Foundation. This support enabled me to spend two productive and enjoyable months at Carnegie Mellon University – Qatar. I am grateful to Renee Barcelona, Eleanore Adiong and Fadhel Annan for making my life easier and my friends Fabiha, Faheem, Rosemary, Rachelle, Adnan, Marjorie, Justin and Muhammed for their support and friendship during my stay.

I would like to thank Tübitak for its financial support throughout my studies.

I am grateful to my parents for their endless love and support. I am indebted to my dear sister Süveyda for her support, friendship and love. I am lucky to have you all in my life.

SYNTAX-TO-MORPHOLOGY ALIGNMENT AND CONSTITUENT REORDERING IN FACTORED PHRASE-BASED STATISTICAL MACHINE TRANSLATION FROM ENGLISH TO TURKISH

Reyyan Yeniterzi

MS Thesis, 2009

Thesis Supervisor: Prof. Dr. Kemal Oflazer

Keywords: Statistical Machine Translation, Factored Translation Model, Syntactic Alignment and Reordering

ABSTRACT

English is a moderately analytic language in which the meaning is conveyed with function words and the order of constituents. On the other hand, Turkish is an agglutinative language with free constituent order. These differences together with the lack of large scale English-Turkish parallel corpora turn Statistical Machine Translation (SMT) between these languages into a challenging problem.

SMT between these two languages, especially from English to Turkish has been worked on for several years. The initial findings [El-Kahlout and Oflazer, 2006] strongly support the idea of representing both Turkish and English at the morpheme-level. Furthermore, several representations and groupings for the morphological structure have been tried on the Turkish side. In contrast to these, this thesis mostly focuses on the experiments on the English side rather than Turkish. In this work we firstly introduce a new way to align the English syntax with the Turkish morphology by associating function words to their related content words. This transformation solely depends on the dependency relations between these words. In addition to this improved alignment, a syntactic reordering is performed to get a more monotonic word alignment. Here, we again use dependencies to identify the sentence constituents and perform reordering between them so that the word order of the source side will be close to the target language.

We report our results with BLEU which is a measure that is widely used by the MT community to report research results. With improvements in the alignment and the ordering, we have increased our BLEU score from a baseline score of 17.08 to 23.78, which is an improvement of 6.7 BLEU points, or about 39% relative.

İNGİLİZCEDEN TÜRKÇEYE FAKTÖRLÜ SÖZCÜK ÖBEĞİ TABANLI İSTATİSTİKSEL BİLGİSAYARLI ÇEVİRİDE SENTAKS-MORFOLOJİ EŞLEŞTİRİLMESİ VE ÖGE YENİDEN SIRALANMASI

Reyyan Yeniterzi

MS Tezi, 2009

Tez Danışmanı: Prof. Dr. Kemal Oflazer

Anahtar Kelimeler: İstatistiksel Bilgisayarlı Çeviri, Faktörlü Çeviri Modeli, Sentaks ile Eşleştirme ve Yeniden Sıralama

Özet

İngilizce, anlamın işlev sözcükleri ve ögelerin dizilimi ile ifade edildiği bir dildir. Türkçe ise serbest öge dizilimi olan, sondan eklemeli bir dildir. Bu farklılıklar büyük çapta bir İngilizce-Türkçe paralel veri eksikliğiyle bir araya gelince, bu diller arasındaki istatistiksel dil çevrisini zorlaştırmaktadır.

Bu iki dil arasında, özellikle İngilizceden Türkçeye, istatistiksel dil çevrimi bir süredir üzerinde çalışılan bir konudur. Bu konuya ilişkin ilk sonuçlar [El-Kahlout and Oflazer, 2006] hem Türkçenin hem de İngilizcenin biçimbilimsel analiz yapılarak ek düzeyinde çalışılmasını destekler tarzdadır. Ayrıca, Türkçe tarafında biçimbilimsel olarak bir takım farklı gösterimler ve gruplamalar da denenmiştir. Bunlara karşılık bu tez Türkçeden daha çok İngilizce tarafındaki deneylere yoğunlaşmaktadır. Bu çalışmada ilk olarak İngilizcedeki işlev sözcükleri, ilgili içerik kelimeleri ile birleştirerek geliştirdiğimiz İngilizce sentaksıyla Türkçe morfolojisi arasında yeni bir eşleştirme yönte mini tanıtıyoruz. İngilizcede yaptığımız bu değişim, yalnızca kelimeler arasındaki bağlılık analizine dayanmaktadır. Bu geliştirilmiş eşleştirmenin yanında, sentaks yönünden yeniden sıralamalar yaparak daha sıralı kelime eşleştirmeleri oluşturmaya çalıştık. Kaynak dilin kelime sırasını hedef dildekine yaklaştırmak için de yine bağlılık analizi kullanarak cümlenin öğelerini teşhis ettik ve yeniden sıralamalar gerçekleştirdik.

Sonuçlarımızı dil çevrimi çalısmalarında çok sık kullanılan BLEU değerlendirme aracı ile elde ettik. Eşleştirme ve sıralamadaki gelişmelerle birlikte BLEU skorumuzu 17.08 den 23.78'e çıkararak 6.7 puanlık bir artış sağladık.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Motivation

Machine Translation (MT) is the application of computers to automatically translate a text or a speech from one language to another. MT is one of the very first applications of computers starting in 40's. Since then, it has been an important topic of research for social, political, commercial and scientific reasons [Arnold et al., 1993], and now in the age of Internet and globalization, the need for MT is more than ever.

Nowadays, international organizations like the United Nations (UN) and the European Union (EU), have to translate their documents to a number of languages. Furthermore, international companies such as Microsoft or IBM are producing documentations and manuals in many languages. Most of these organizations and companies use human translators to deal with this translation issue; however, since manual translation is a labor and time intensive task and there are never enough translators, this solution becomes an expensive one. These reasons motivate researchers to work on efficient MT systems with good output quality.

Another motivation for the MT research has been the rapid increase in the popularity of the Internet. Within the last decade, the Internet has become the ultimate source of information. Everyday, millions of people use search engines to find the desired information on the web. However, most of the time users cannot exploit the information found since it is in a different language. Several search engines such as Google Translator, Yahoo! Babel Fish, use translation systems to give their users a better

search experience. These systems help the reader to understand the general content of the foreign language text, but unfortunately they do not always produce perfect or even accurate translations. Therefore, there is still a lot of room for improvement and this motivates the researchers to focus on improving the current methods and developing new ways to produce high quality MT systems.

Currently, the state-of-the-art approach in MT research is the Statistical Machine Translation (SMT) method, which was proposed by IBM in 1990s. SMT is a statistical approach for MT which derives its model from the analysis of bilingual parallel sentences. It is completely an automatic method which does not require any manual translation rules or specific tailoring for any specific language. Because of these reasons, it is by far the most widely used machine translation method in MT community.

In this thesis, we use a certain novel SMT approach to translate from English to Turkish. This approach introduces a new method to align syntax and morphology by associating function words to their dependent content words. We also experiment with syntactic reordering between sentence constituents to see if better translation can be obtained with close word order.

1.2 Outline

The organization of this thesis is as follows: Chapter 2 starts with an introduction to MT then continues with an overview of SMT and SMT from English to Turkish. Chapter 3 describes the syntax-to-morphology alignment by explaining transformation procedures and giving detailed examples. In Chapter 4, we present our experiments with syntactic reordering. Finally, in Chapter 5 we conclude with a summary of the thesis.

Chapter 2

STATISTICAL MACHINE TRANSLATION

2.1 Introduction to Machine Translation

Machine Translation is the automatic translation of a *source text* into another language, which is referred as the *target language*, while keeping the meaning same. This translation process has three main steps which are (1) analysis of the source text into a certain representation, (2) transforming this representation and (3) generating a text in the target language from this representation. These three steps require an extensive knowledge of the vocabulary, syntax and semantics of both languages. Acquiring and using this knowledge correctly is the main challenge of MT.

2.1.1 Challenges in MT

MT is a challenging problem because of the ambiguity and differences between languages. In order to develop a high quality MT system, we have to know about these challenges and act accordingly.

Languages contain ambiguity at all levels, and this is a problem for almost all natural language processing applications. So, ambiguity also complicates the analysis step of MT. For instance, a sentence like "I saw a woman with a telescope." can be interpreted in two different ways: whether (1) the action of seeing is performed with a telescope or (2) the woman has a telescope. Furthermore, word sense ambiguity may also cause different interpretations. The word "tear" in the sentence "There is a tear

on her shirt." can mean either (1) a damage or (2) a fluid flowing from the eye as a result of emotion. In order to get a correct translation, such semantic ambiguities have to be resolved in the analysis step.

Another challenge in MT is the lexical or syntactic differences between source and target languages. In terms of lexical differences, an interesting problem is the lexical gap: no word or phrase in the target language can express the meaning of a word in the source language. For example in Turkish, the word "bacanak", the husband of one's wife's sister, does not have any direct translation in English. Furthermore, there is also the problem of a word having multiple meanings such as our previous example "tear".

An additional language divergence, which complicates MT, is the syntactic differences between target and source language. A common example to this is the different constituent structures of languages. Most of the languages such as English, French and German have Subject-Verb-Object (SVO) constituent order. On the other hand there are languages, like Turkish, which have Subject-Object-Verb (SOV) order. In addition to this top level structural difference between languages, there are some other syntactic variations, such as verb argument changes or differences in passive constructions [Lavie, 2008] between languages. Currently these differences are the main challenge in MT and they have to be tackled in order to develop high quality systems.

2.1.2 Approaches to MT

Approaches to MT make use of the three steps that we have mentioned before: Analysis, Transfer and Generation. These steps and their relations to the source and target texts are represented in the Vauquois triangle in Figure 2.1. This triangle shows the depths of the intermediate representation and the most common approaches used in MT.

At the bottom of the triangle we see the simplest approach which is *direct translation*. This approach does not produce any intermediate representation, but it relies on some shallow analyses (e.g., morphological analysis) in the translation. Direct translation also uses some reordering rules in order to do local word order adjustments. This approach is usually easy to implement and can produce translations that can give a rough idea about the source content.

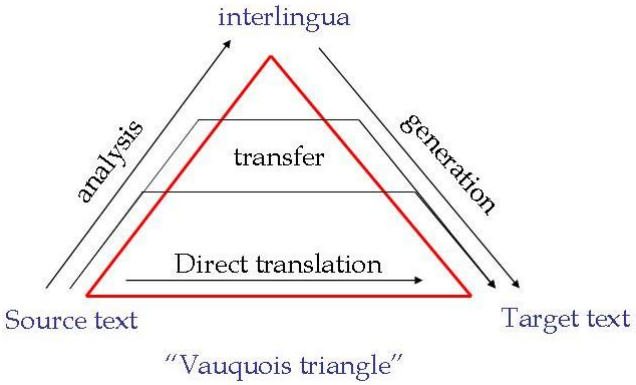


Figure 2.1: Vauquois MT triangle

When we go higher in the triangle, the methods employ deeper analyses such as syntactic and semantic analyses. In syntactic analysis, the source sentence is parsed to produce a parse tree. Then, this source language structure is transferred into the target language structure by applying sets of linguistic rules to transform trees. Finally, the surface sentence is generated in target language from the transformed tree. This transfer approach requires parsers and generators for each language pair which require substantial manual labor.

At the top of the triangle we see the *interlingua approach*, which relies on a "language independent representation". In this approach, the source text is analyzed into the symbolic representation of its "meaning". Then without any transformation, this representation is used to generate the target text. This approach has both advantages and disadvantages. In multilingual MT systems, it gives the advantage of not developing transfer rules for each language pair. On the other hand, developing a language independent representation for a wide domain is extremely difficult.

Most of these approaches are rule-based methods which rely on building linguistically grounded rules and bilingual dictionaries. Therefore, creating these systems are both expensive and labor intensive. In 1990's with the availability of parallel corpora, researchers started to work on statistical approaches. In the next section, we are going to describe these statistical MT approaches in detail.

2.2 Overview of Statistical Machine Translation

Statistical Machine Translation (SMT) approach uses statistical models to find the most probable target sentence (t) given the source sentence (s). Mathematically speaking, we can represent this as follows;

$$\hat{t} = \arg\max_{t} P(t \mid s) \tag{2.1}$$

where t ranges over all possible target sentences. Applying Bayes' theorem to Equation 2.1 gives us

$$\hat{t} = \arg\max_{t} P(s|t)P(t)/P(s) \tag{2.2}$$

In this equation, $P(s)$ is constant for every possible t, so we can ignore it and get

$$\hat{t} = \arg\max_{t} P(s|t)P(t) \tag{2.3}$$

Equation 2.3 can be interpreted in the following way: The most probable target sentence $\hat{t}$ is that t which maximizes the product of $P(s|t)$ and $P(t)$. Here $P(s|t)$ is called the *translation model* which is the probability of s being the translation of t. The other factor $P(t)$ is called the *language model* and it is the probability of t being a valid sentence in the target language.

A typical SMT system uses these two models and a decoder to search and find the most probable translation. An overview of this SMT process is presented in Figure 2.2. The translation model is generated from the bilingual texts, while the language model is estimated from the target text only. The decoder uses these two models and searches through the space of possible translations to identify the most probable one. We are now going to describe these three components of SMT in detail.

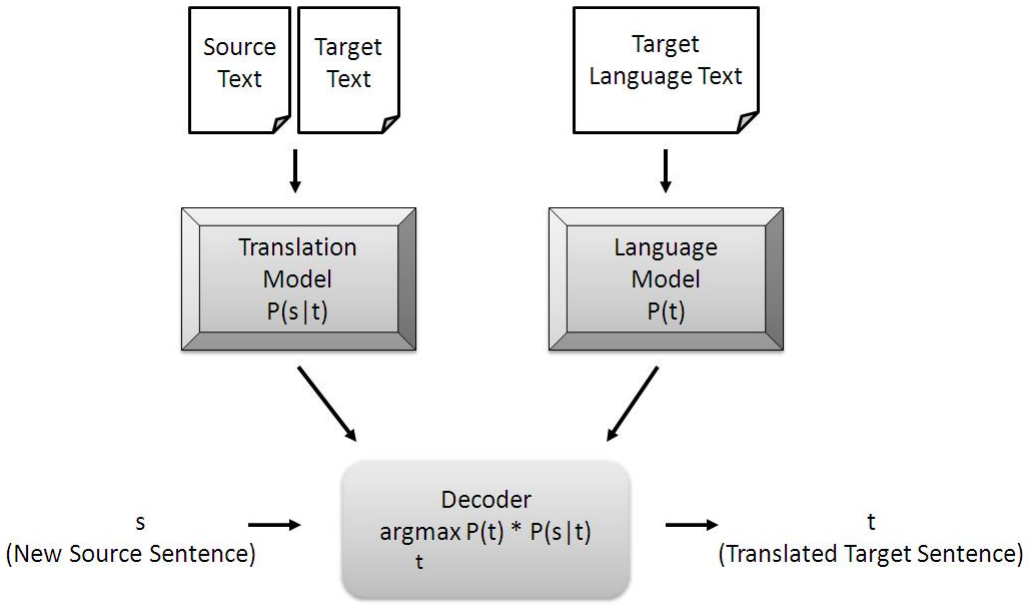


Figure 2.2: Overview of SMT

2.2.1 The Components of a SMT System

2.2.1.1 Language Model

The language model (LM), is a statistical model that can assign probabilities to sequences of words in a language: more likely or grammatical word sequences get high probabilities while word salads or ungrammatical sequences get very low probabilities. This component is used to ensure that words are in right order so that the sentence is syntactically correct and fluent. In a LM, the probability of seeing a sentence t of $w_1...w_n$ is modeled as following:

$$P(t) = P(w_1)P(w_2|w_1)P(w_3|w_1\ w_2)...P(w_n|w_1\ w_2\ \ldots\ w_{n-1}) \tag{2.4}$$

In the equation above, $P(w_1)$ is the probability of seeing w_1 independently, $P(w_2|w_1)$ is the probability of seeing w_2 after w_1, $P(w_3|w_1\ w_2)$ is the probability of seeing w_3 after the $w_1\ w_2$ phrase and $P(w_n|w_1\ w_2\ \ldots\ w_{n-1})$ is the probability of seeing the last word w_n after seeing all $n-1$ preceding words. The product of all these probabilities gives us the probability of seeing that sentence, via the chain rule.

For a given word, looking at all the preceding words in the sentence is not very realistic due to sparseness issues. A practical approach is to assume a *Markov process* so that a word is conditioned by a small number of past neighbors. If all words in a model depend on the preceding $n-1$ words, then that model is called an n-gram word model [Manning and Schütze, 1999]. Currently, 3-gram (trigram) or 4-gram models are the mostly used models in SMT. An example probability calculation of a trigram model of a sentence is given below.[1]

$$
\begin{aligned}
P(Tourists\ are\ very\ fond\ of\ Turkish\ hospitality) \;=\; & P(Tourists| < s > < s >) * \\
& P(are|Tourists\ < s >) * \\
& P(very|Tourists\ are) * \\
& P(fond|are\ very) * \\
& P(of|very\ fond) * \\
& P(Turkish|fond\ of) * \\
& P(hospitality|of\ Turkish) * \\
& P(< /s > |Turkish\ hospitality) * \\
& P(< /s > |hospitality\ < /s >)
\end{aligned}
$$

Trigram probabilities are estimated via counts in the corpus.
e.g.

$$P(w_3|w_1\ w_2) \cong count(w_1\ w_2\ w_3)/count(w_1\ w_2) \tag{2.5}$$

If a model is estimated from a small amount of data, then many n-grams may not exist in the model and therefore their probability will be equal to zero. Various smoothing methods exist to alleviate this problem [Manning and Schütze, 1999]

Currently there are several publicly available LM tools. The most popular is the SRI LM Toolkit [Stolcke, 2002] which has been initially developed for speech recognition.

[1] $< s >$ indicates the start of a sentence and $< /s >$ represents the end of a sentence

Other similar tools that are used by MT community are the IRSTLM tool [Federico et al., 2008] and the CMU/Cambridge LM Toolkit [Clarkson and Rosenfeld, 1997].

2.2.1.2 Translation Model

The translation model $P(s|t)$ captures the probability of sentence s being the translation of sentence t. It is estimated from a bilingual parallel corpus. Since computing this probability at the sentence level is almost impossible, words and their alignments are used instead [Brown et al., 1993]. This model, usually known as IBM Model 3, allows one-to-many word alignments which is represented with vector a. These alignment probabilities of words are used to calculate the $P(s|t)$.

$$P(s|t) = \sum_{a} P(a, s|t) \tag{2.6}$$

Given a sentence t, the probability of producing a particular sentence s and an alignment a between s and t is the product of several other probabilities. These are

- Translation Probability : $t(s_j|t_i)$ is the probability of word t_i being translated into word s_j.
- Fertility Probability : $n(\phi_i|t_i)$ is the probability of translating t_i into ϕ_i number of words.
- Distortion Probability : $d(j|i, l, m)$ is the probability of aligning the target word in position i with the source word in position j given the sentence lenghts l and m.

where m is the number of words in sentence s, l is the number of words in sentence t, s_j is the source word in position j, t_i is the target word in position i, ϕ_i is the fertility of word in position i.

These probabilities are estimated using the Expectation Maximization (EM) algorithm. This algorithm starts with some initial random estimate of the parameters and uses these parameters to compute the probability of alignments. Then these parameters

are re-estimated by collecting counts. These steps are repeated until the parameters converge. [Jurafksy and Martin, 2000]

After training the Language and Translation Models, SMT system is ready to decode new sentences.

2.2.2 Decoding

The main task of this step is to search and find the most probable target sentence given the source sentence and the already trained models. Each potential translation output is called a *hypothesis*. There are infinitely many potential target sentences and so decoding is known to be an NP-complete problem [Knight, 1999]. In order to find the best translation effectively within this large search space, several heuristic search algorithms have been developed. One efficient commonly used method is the beam search. The idea behind this approach is to keep hypotheses in stacks based on their number of translated words. If an hypothesis is extended by translating more words then it has to be moved to the corresponding stack. Later, if necessary, that stack is pruned by removing the least probable hypothesis.

2.3 Phrase-Based Statistical Machine Translation

In previous section, we summarized word-based SMT systems, in which the translations are performed with word-by-word mappings. These models can do one-to-many alignments but not many-to-one. To overcome this limitation, *phrase-based* SMT systems have been developed, which can handle many-to-many translations. Another advantage of phrase-based systems is that since they use any sequence of words, they can encapsulate the local context and the local reordering.

Phrase translations can be learned by several ways. One method is to use the *alignment templates* [Och et al., 1999]. This method starts with training word alignment models and then uses both Viterbi paths to extract phrases. An improved method was suggested by Koehn *et all.* [Koehn et al., 2003]. In this approach, the parallel corpus is aligned bidirectionally in order to generate two word alignments. Starting from the

intersection of these alignments, new alignment points which exist in the union and connect at least one previously unaligned word are added. The algorithm starts with the first word and continues adding new alignment points from the rest of the words in order. With this method all aligned phrase pairs that are consistent with the word alignment are collected. Finally, the probabilities are assigned to these phrase pairs by doing relative frequency calculations.

2.3.1 Factored Translation Models

Currently, the phrase-based translation approach is the most promising state-of-the-art approach in SMT, but still it does not use any linguistic information such as morphology or syntax. In order to integrate these additional annotations to the word level, an extension *factored translation* has been developed [Koehn and Hoang, 2007]. This model does not just represent the word itself but also contain some other annotations like lemma, part-of-speech (POS), morphology as shown in Figure 2.3. Each of these annotations is called a *factor*.

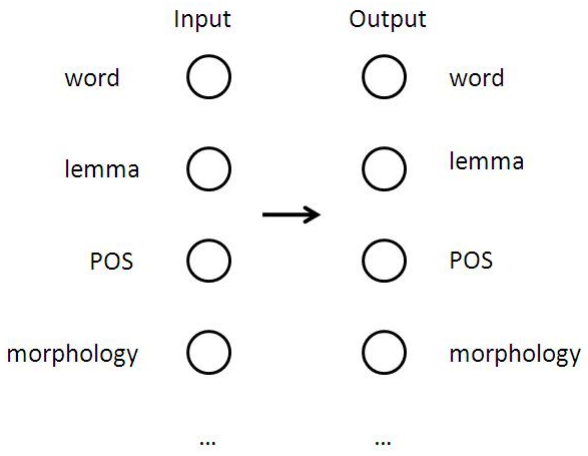


Figure 2.3: Factored representations of input and output words

Factored translation models are meant to be used for morphologically rich languages. In morphologically rich languages, different word forms are derived from the same lemma which results in poor statistics when limited training data is used. In situations like these, factored translation gives us a more general approach which translates

lemma, and morphology separately and then generates the target surface form. Such a model is illustrated in Figure 2.4.

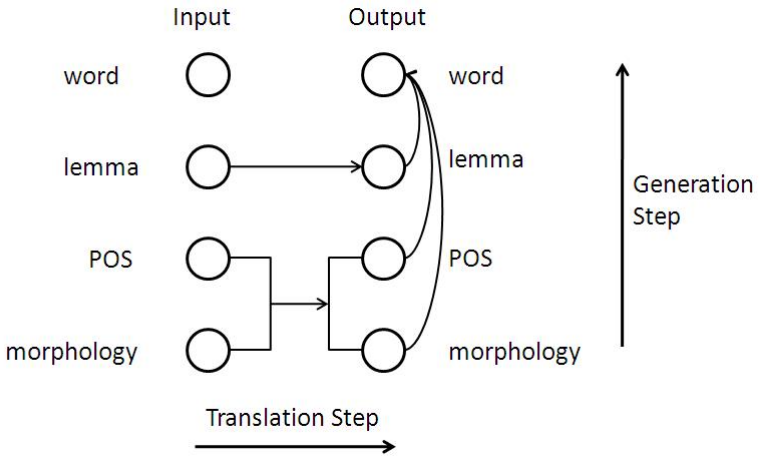


Figure 2.4: An example factored model for morphologically rich languages

In Figure 2.4, the arrows represent the mapping steps. There are two kinds of mapping steps. The first one is the translation step which maps input factors to output factors at the phrase level. Translation steps are represented with the horizontal arrows in Figure 2.4. There are two translation steps in this model; (1) translation of input lemmas to output lemmas and (2) translation of input part-of-speech (POS) and morphology to output POS and morphology.

The other mapping step is called the generation step. This step is used to map output factors into other output factors at the word level. In Figure 2.4, this step is represented with the curved vertical lines, which describe the generation of surface form from lemma, POS and morphology.

While training the factored translation models, the same methods are used to learn the phrase tables from word-aligned parallel corpora. On the other hand the generation tables are learned from just the target side of the parallel corpus by using word level frequencies. Similarly, in factored model decoding instead of just using one phrase table, we use multiple phrase tables and generation tables.

2.4 Evaluation of SMT Outputs

Last but not least, there is the task of evaluating the translation quality. There are some manual approaches for this task which are performed by human experts. One of them is the SSER (Subjective Sentence Error Rate), in which the translations are classified according to their quality ranging from 0 to 10 [Niessen et al., 2000]. In order to deal with the subjective nature of this approach, these evaluations have to be performed by several people. Therefore this approach is expensive, labour intensive and time consuming.

Since MT researchers need instant feedback about their work and improvements, several automatic approaches to MT evaluation have been proposed. These score metrics and tools are developed with the aim of returning a score which is in strong correlation with the human evaluator.

Among those tools, BLEU (Bilingual Evaluation Understudy) [Papineni et al., 2001] is the most widely used one. BLEU is a n-gram-based evaluation metric which makes sure that a good candidate has similar word choice and order with the reference sentence. Moreover, BLEU uses a modified version of n-gram precision to penalize repetitions in a sentence and the authors introduced a brevity penalty for candidate sentences that are shorter than the reference.

BLEU is a language independent tool and it is used widely by the MT community to report performance results. BLEU returns a score between 0 to 1. A score close to 1 indicates that the candidate is really similar to the reference, therefore it is a good translation.

2.5 SMT from English to Turkish

SMT from English to Turkish is a challenging problem due to the morphological and grammatical distance between these languages. While English has a limited morphology, Turkish is an agglutinative language with a very rich morphological structure. In terms of the constituent order, English is rather strict on using Subject-Verb-Object

order, while Turkish uses a more flexible order which is mostly Subject-Object-Verb. These differences together with some other practical problems make SMT from English to Turkish a difficult problem.

2.5.1 Challenges

Like most other statistical applications, SMT is a data driven approach. Its success mostly depends on the amount and the quality of the bilingual parallel texts. Currently, this seems to be a significant problem for the English-Turkish pair. In this thesis we work with approximately 50K sentences, while a good SMT system requires at least a few million parallel sentences. Although the number of sentences in this parallel corpus can be increased by using web and some other resources, it requires a significant collection and cleanup process. Therefore, we don't think this problem will be resolved in the near future, for the Turkish-English language pair.

Another challenge of SMT from English to Turkish arises from the rich inflectional and derivational morphology of Turkish. In Turkish a single word may contain many morphemes and each of these represents a different grammatical meaning. In word level alignment, this results in the alignment of one Turkish word with a phrase of words on the English side. For instance, the Turkish word 'tatlandırabileceksek' is translated into a phrase like 'if we are going to be able to make [something] acquire flavor' [Oflazer, 2008]. Another issue that is caused by the rich morphology of Turkish is the translation of very frequent English words into words with very low frequency in Turkish side. An example to this is given by El-Kahlout and Oflazer over the root word *faaliyet* 'activity' [El-Kahlout and Oflazer, 2006]. They showed that for 41 occurrences of the word 'activity' (singular and plural), there are only 14 different forms of *faaliyet*, such as faaliyetlerinde (in their activities), faaliyetlerin (of the activities), etc., to which it is aligned. To overcome these alignment and sparseness problems, a morphological analysis is performed on both Turkish and English texts.

The word order variations between English and Turkish may also be a problematic issue. In addition to the top level word order difference, there are also ordering differences in subordinate clauses, passive voices and phrases. These word order differences

result in a larger search space in decoding step, which will increase the translation time. In order to deal with this problem, some reordering techniques can be tried which will produce more monotonic alignments.

2.5.2 Previous Work

First research on MT from English to Turkish has started in early 1980s as a master's thesis [Sagay, 1981], which much later was developed into an interactive machine translation environment called *Çevirmen*. After this first system, two other approaches have been tried in late 1990s. One of them used structural mapping in a transfer-based approach [Turhan, 1997] and other one developed a prototype English-to-Turkish interlingua-based machine translation system by using KANT knowledge-based MT system [Hakkani-Tür et al., 1998].

Recently, several statistical approaches have been tried with English-Turkish pair. Türe proposed a Hybrid Machine Translation System from Turkish to English [Türe, 2008]. Moreover, Oflazer and El-Kahlout developed a prototype English-Turkish SMT system by exploring different representational units of Turkish morphology [Oflazer and El-Kahlout, 2007, El-Kahlout, 2009].

Chapter 3

SYNTAX TO MORPHOLOGY ALIGNMENT

3.1 Motivation

English is a moderately analytic language [Barber, 1999] in which grammatical relations are expressed by words instead of morphemes. These words such as prepositions, pronouns, auxiliary words, articles, which have very little lexical meaning are called function words. There are also content words which represent the lexical items. These words include nouns, verbs, adjectives and adverbs. English grammar mostly describes the syntactic relationship between these two groups of words rather than their morphology. This however doesn't hold for the Turkish grammar. As we mentioned in Section 2.5, Turkish is an agglutinative language in which words are made up of joining morphemes together. Each of these morphemes represents one grammatical meaning. Furthermore agglutinative languages tend to have high number of morphemes per word. Thus, in Turkish, most of the grammatical relations are determined by morphological features.

These differences between English and Turkish complicate the word alignment and result in the alignment of one Turkish word with a bunch of English words as in the example given in Section 2.5.1. In this thesis, we propose a method to align English syntax with Turkish morphology via a preprocessing step on the English side so that the English sentences look more like Turkish.

3.1.1 Overview of the Approach

Machine translation between syntactically similar languages is usually of better quality than between languages that are not so close [Hajič et al., 2000]. With this observation in mind, our approach focuses on decreasing the structural gap between English and Turkish sentences. This can be done by performing syntactic transformations and word reorderings. Our overall approach covers both of these, but we will talk more about the transformations in this chapter and leave the discussion on reordering to the next chapter.

Since we are translating from English to Turkish, we also develop transformation methods from English to Turkish so that the structure of English sentences will become similar to the Turkish sentences. As we have shown before, function words of English sentence usually become morphemes when they are translated into Turkish. We perform this change as a preprocessing step and append these function words to their related content words before giving them to the SMT system. The relationships between these words are found by using syntactic analysis.

Our approach starts with some analysis on both Turkish and English sentences. We perform a morphological analysis on Turkish sentences [Oflazer, 1993] and a part-of-speech tagging on English corpus [Toutanova et al., 2003]. Then we give our tagged English corpus to a dependency parser [Nivre et al., 2007] to find the dependency relations. After all these analyses, we apply the transformation rules depending on the relations and finally give our parallel corpus to training.

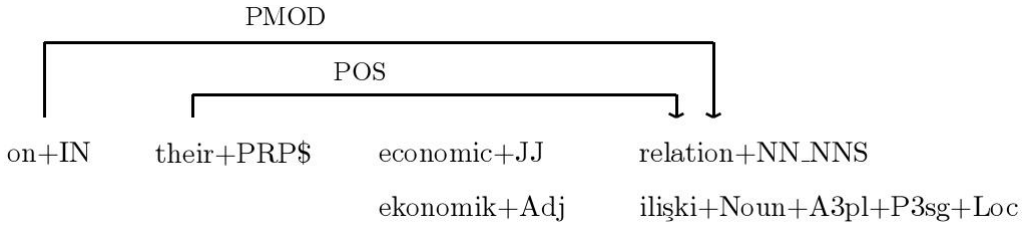
3.1.2 Examples

Before going into the implementation details, we summarize our approach over some examples. For instance let's assume we are given the below aligned pair.

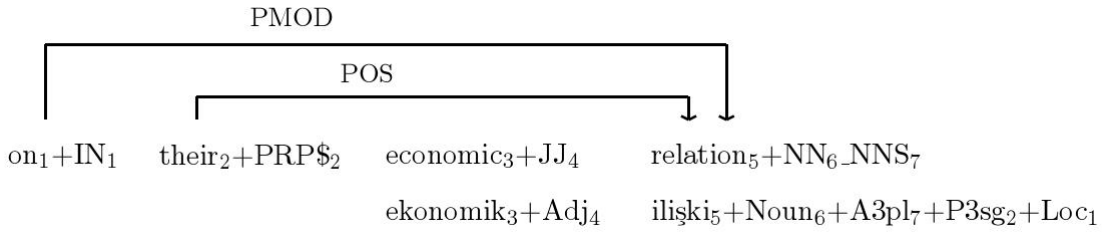
on	their	economic	relations
		ekonomik	ilişkilerinde

As it is seen above, the function words *on* and *their* are not aligned with any of the Turkish words, but the content words are aligned with their Turkish translations. If we

tag and parse the English sentence and give the Turkish sentence to a morphological analyzer, we will get the following representations.[1]



Here one can see the POS tags and morphemes of the words, and the dependencies between words. From the labels on the dependency arrows, it is understood that *on* is the preposition modifier and *their* is the possessive of the word *relations*. If we align all these lemmas, tags and morphemes with each other by using coindexation, we will get something like



Here we see that English lemmas are aligned with Turkish lemmas (3, 5), English POS tags are aligned with Turkish POS tags (4, 6) and an English morpheme is aligned with a Turkish morpheme (7). Furthermore English function words should be aligned with the rest of the Turkish morphemes (1, 2); because *on+IN* becomes the *+Loc* morpheme and *their+PRP$* becomes the *+P3sg* morpheme on the Turkish side. When we perform

[1]The meanings of the tags are as follows:

Dependency Labels	
PMOD	Preposition Modifier
POS	Possessive
Tags in English Sentence	
+IN	Preposition
+PRP$	Possessive Pronoun
+JJ	Adjective
+NN	Noun
+NNS	Plural Noun
Tags in Turkish Sentence	
+A3pl	3rd person plural possessive
+P3sg	3rd person singular possessive
+Loc	Locative case

our transformations and append those function words to the related content word, our sentences will become

economic$_3$+JJ$_4$	relation$_5$+NN$_6$_NNS$_7$_their$_2$+PRP\$$_2$_on$_1$+IN$_1$
ekonomik$_3$+Adj$_4$	ilişki$_5$+Noun$_6$+A3pl$_7$+P3sg$_2$+Loc$_1$

As it is seen from the example, these transformations are performing syntax to morphology alignments and capturing English syntax as complex tags on appropriate head words. Since we perform these transformations in a specific order, a unique word is produced at the end of transformations. For the same combination of transformations, same order is applied to all words.

In the rest of this thesis, we will represent these transformations in three steps, as shown in Figure 3.1. Here the first step shows the word level alignments of the original sentences in their surface forms. The second step presents the sentences after the analyses are performed. This representation also includes the alignments of smaller components. The last step is the output sentence after the transformations are completed.

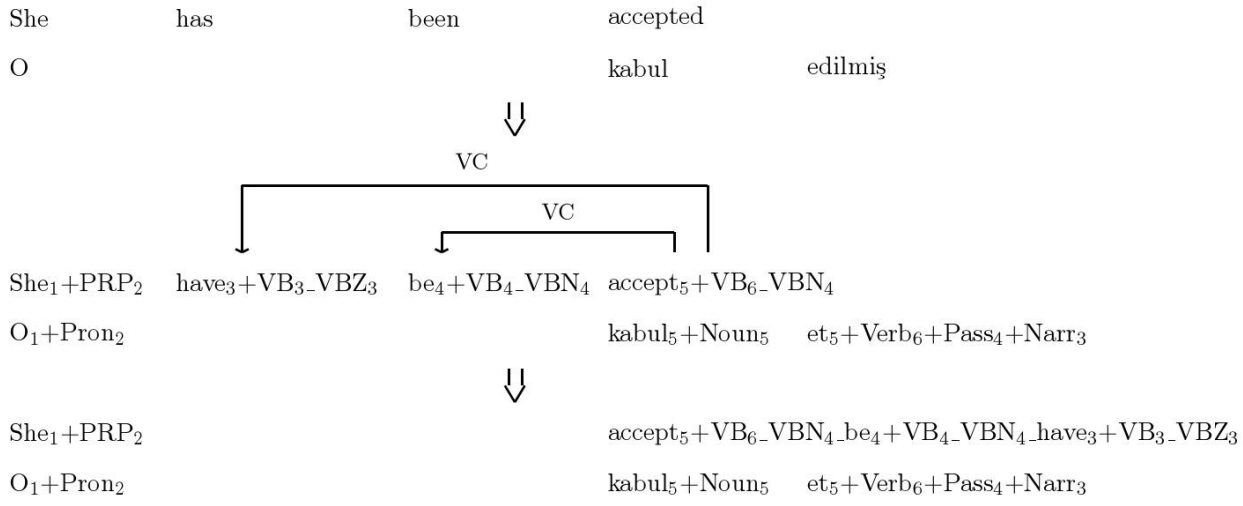


Figure 3.1: An example for transformation step

3.2 Implementation

3.2.1 Data Preparation

We worked on an English-Turkish parallel corpus which is a collection of European Union documents, decisions of the European Court of Human Rights and several treaty texts. This data consists of approximately 50K sentences with an average of 23 words in English sentences and 18 words in Turkish sentences.

With the aim of understanding these texts better both syntactically and semantically, we perform several analyses. For the English side, we start with part-of-speech tagging and then continue with parsing. On the Turkish side, we perform a morphological analysis and morphological disambiguation. In this section we will give more details about each of these steps.

3.2.1.1 Tagging

Part-of-speech (POS) tagging is the process of assigning part-of-speech tags, such as noun, verb, adjective and adverb, to words depending on the word itself and the context. We apply Stanford Log-Linear Part-of-Speech Tagger [Toutanova et al., 2003] which outperforms most of the other taggers by making use of bidirectional inference and the broad use of lexicalization with suitable regularization. We use the already trained model for English that comes with the tagger. In addition to this we also use TreeTagger in order to find the lemmas of words [Schmid, 1994]. Both of these tools use the Penn Treebank English POS tag set [Marcus et al., 1994]. An example output after tagging is given below.

```
The+DT initiation+NN of+IN negotiation+NN_NNS will+MD
represent+VB the+DT beginning+NN of+IN a+DT next+JJ
phase+NN in+IN the+DT process+NN of+IN accession+NN.
```

3.2.1.2 Parsing

After tagging the English data, we continue with parsing the tagged sentence to extract its grammatical structure. For parsing the English data set, we use the MaltParser [Nivre et al., 2007] with the pretrained model on English [Hall et al., 2008].

An example output of the MaltParser is shown in Figure A.2. As it is seen, there are several fields in the output. These are in order from left to right: token id, word form, lemma, coarse-grained part-of-speech tag, fine-grained part-of-speech tag, head of the current token and the dependency relation of current token with its head [Buchholz and Marsi, 2006].

```
1   the           the           DT   DT   _  2   NMOD
2   initiation    initiation    NN   NN   _  5   SBJ
3   of            of            IN   IN   _  2   NMOD
4   negotiations  negotiation   NNS  NNS  _  3   PMOD
5   will          will          MD   MD   _  0   ROOT
6   represent     represent     VB   VB   _  5   VC
7   the           the           DT   DT   _  8   NMOD
8   beginning     beginning     NN   NN   _  6   OBJ
9   of            of            IN   IN   _  8   NMOD
10  a             a             DT   DT   _  12  NMOD
11  next          next          JJ   JJ   _  12  NMOD
12  phase         phase         NN   NN   _  9   PMOD
13  in            in            IN   IN   _  12  ADV
14  the           the           DT   DT   _  15  NMOD
15  process       process       NN   NN   _  13  PMOD
16  of            of            IN   IN   _  15  NMOD
17  accession     accession     NN   NN   _  16  PMOD
```

Figure 3.2: An example output of MaltParser

In Figure A.2, `initiation` is the subject of the modal `will` which is the root or the head of the sentence. `beginning` is the object of the sentence while the phrase starting with `in` is the adverb. Furthermore, there are several noun modifiers (NMOD) and preposition modifiers (PMOD) which are used to link these words with each other.

3.2.1.3 Morphological Analysis

On the Turkish side, to get more insight on the internal structure of sentence and words, we have to look at the morphemes. Since morphemes contain most of the necessary

grammatical information, we perform a morphological analysis and extract the morphological features of each word. We use a Turkish morphological analyzer [Oflazer, 1993], which basically segments the morphemes and then normalizes the lemma if it has been modified because of the morphemes and maps morphemes to features. An example input and output sentence can be

```
Müzakerelerin başlaması , katılım sürecinin bir sonraki
aşamasının başlangıcını temsil edecektir
```

⇓

```
müzakere+Noun+A3pl+Gen
başla+Verb+Inf2+P3sg
,+Punc
katılım+Noun
süreç+Noun+P3sg+Gen
bir+Num sonra+Noun+Rel
aşama+Noun+P3sg+Gen
başlangıç+Noun+P3sg+Acc
temsil+Noun
et+Verb+Fut+Cop
```

In the output, each marker with a preceding `+` is a morphological feature. The first marker is the part-of-speech tag of the lemma and the remainder are the inflectional and derivation markers of the word. For example, the word `müzakere+Noun+A3pl+Gen` represents the lemma `müzakere`, which is a `Noun`, with third person plural agreement `A3pl` and genitive case `Gen`.

3.3 Transformations

In this section we describe the transformations that are performed on the English and Turkish sentences in order to close the structural gap between these sentences.

3.3.1 English

On the English side, we use the dependencies between words while doing the transformations. The dependent function words of a content word in English are very much similar to the morphemes of the corresponding Turkish word. In Turkish all the morphemes are suffixes, which means that they are concatenated to the word from the end. To have a similar representation, we also perform the transformations in that way. We place the function word after the content word with an underscore between them. An example sentence before and after the transformation is given below.

```
The+DT initiation+NN of+IN negotiation+NN_NNS will+MD
represent+VB the+DT beginning+NN of+IN a+DT next+JJ
phase+NN in+IN the+DT process+NN of+IN accession+NN
```

⇓

```
initiation+NN_the+DT negotiation+NN_NNS_of+IN
represent+VB_will+MD beginning+NN_the+DT next+JJ
phase+NN_of+IN_a+DT process+NN_in+IN_the+DT
accession+NN_of+IN
```

The following are the detailed descriptions of each of these transformations with examples.

3.3.1.1 Prepositions

A preposition is a function word which puts object noun phrase in a certain relationship with another word: for example in "on my table", "on" is the preposition and "my table" is the object of the preposition. In English, a preposition precedes the noun phrase. On Turkish side, these prepositions are mostly represented with case morphemes that are bound to the related content word. Some of the most commonly used prepositions and corresponding case morphemes are given in Table 3.1.

In the dependency parser output, these prepositions are linked to their object heads with the Preposition Modifier (PMOD) tag. We use these tags to find the prepositions and their related content words and then perform the transformations. Example preposition transformations are given in Figures 3.3 and 3.4.

Turkish	English
+Dat (Dative)	to
+Abl (Ablative)	from
+Loc (Locative)	on, in, at
+Gen (Genitive)	of
+Ins (Instrumental)	with

Table 3.1: Example case morphemes and prepositions

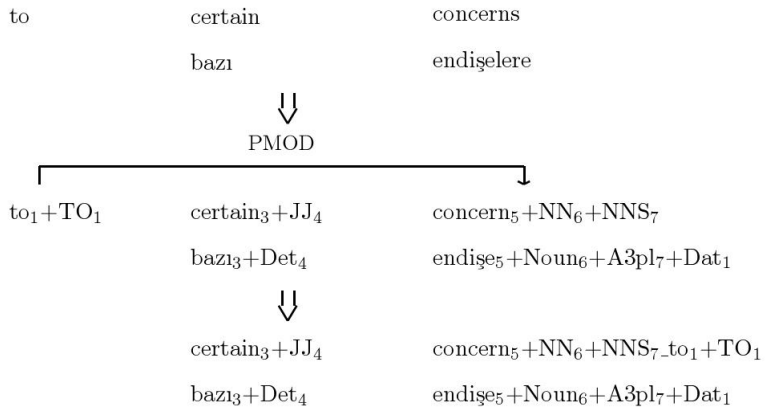


Figure 3.3: An example for preposition transformation

3.3.1.2 Possessives

Possessive Pronouns

Possessive pronouns in English are function words which denote the "possession" of nouns. In English, they precede the word they are specifying, but in Turkish, they are attached to the end of the word as so called possessive suffixes, in addition to being explicitly present as in English. Possessive pronouns of English and Turkish are given in Table 3.2.

Possessor	Turkish	English
1. singular	benim	my
2. singular	senin	your
3. singular	onun	his, her, its
1. plural	bizim	our
2. plural	sizin	your
3. plural	onların	their

Table 3.2: Possessive pronouns in Turkish and English

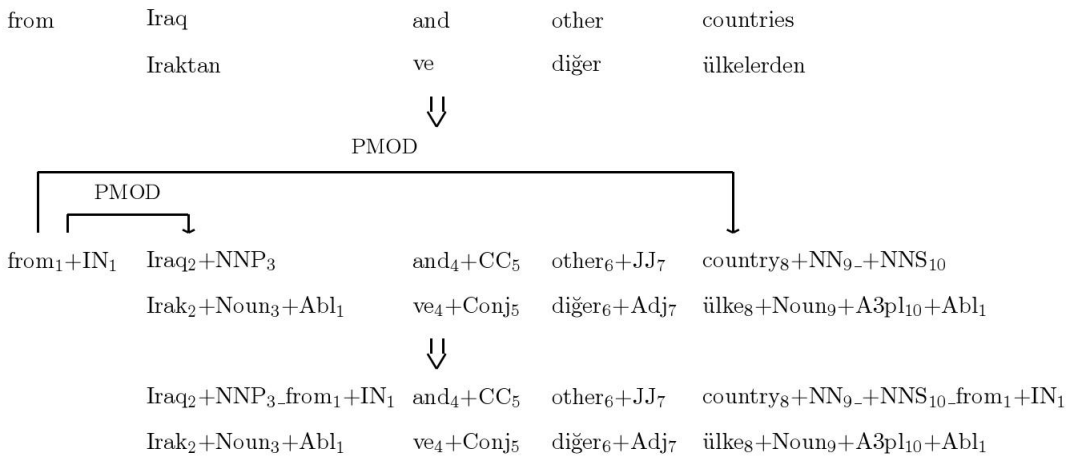


Figure 3.4: An example for preposition transformation

In English, the dependency between a possessive pronoun and a noun is represented with a Noun Modifier (NMOD) label. An example of the possessive pronoun and the related transformation is given in Figure 3.5.

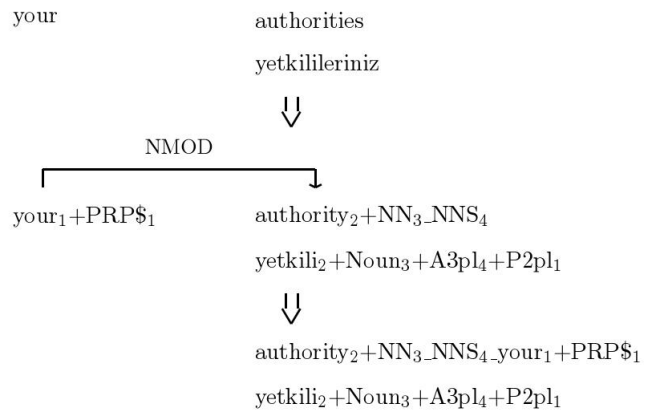


Figure 3.5: An example for possessive pronoun transformation

Possessive Marker

The possessive marker is used to indicate a possession relationship. In Turkish *+Gen* case marking is used to represent this relation. Similarly English uses a morpheme for this grammatical relation instead of a function word. In English, to indicate a possession, *'s* morpheme is suffixed to the noun that is the "possessor". Before the tagging step we separate this suffix and treat it as an individual token. During the parsing step this token becomes a Noun Modifier and in the transformation step it is

again connected to its head noun which is the owner. An example transformation is given in Figure 3.6.

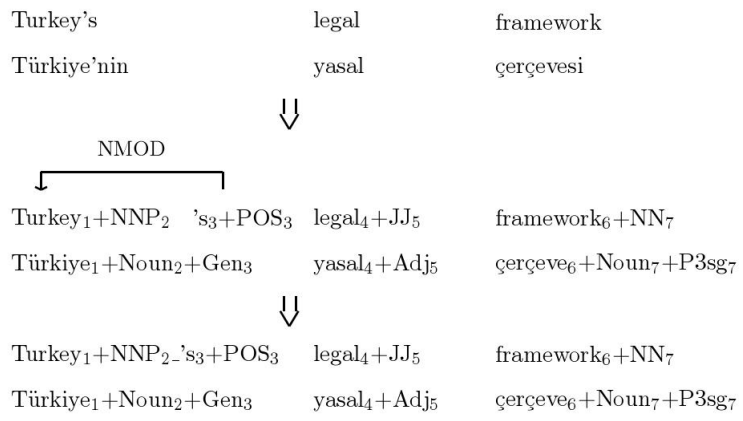


Figure 3.6: An example for possessive marker transformation

3.3.1.3 The Copula "be"

A copula is a verb which links a subject to a predicate which is either a noun phrase or an adjective. In English, the main copular verb is *be*, however some other verbs like *get*, *seem* and *feel* can also be used as copula verbs. Among those verbs, we only focus on *be*.

The copula *be* is used with a predicate noun to describe the subject, or it can be used with a predicate adjective to give an attribute of the subject. In Turkish, both nouns and adjectives can get the *+Cop* morpheme, to become the predicate of the sentence. We apply transformation to both of these part-of-speech when they are used together with the copula *be*. An example for each of them is given in Figures 3.7 and 3.8.

3.3.1.4 Articles

English has three articles. These are *the*, which is the only definite article, and the indefinite articles *a* and *an*. These articles are used together with nouns to indicate whether a reference is specific or general. In Turkish there is no morpheme that is a counterpart to "the", but since they are function words which modifies the content word we also append these articles to the head word.

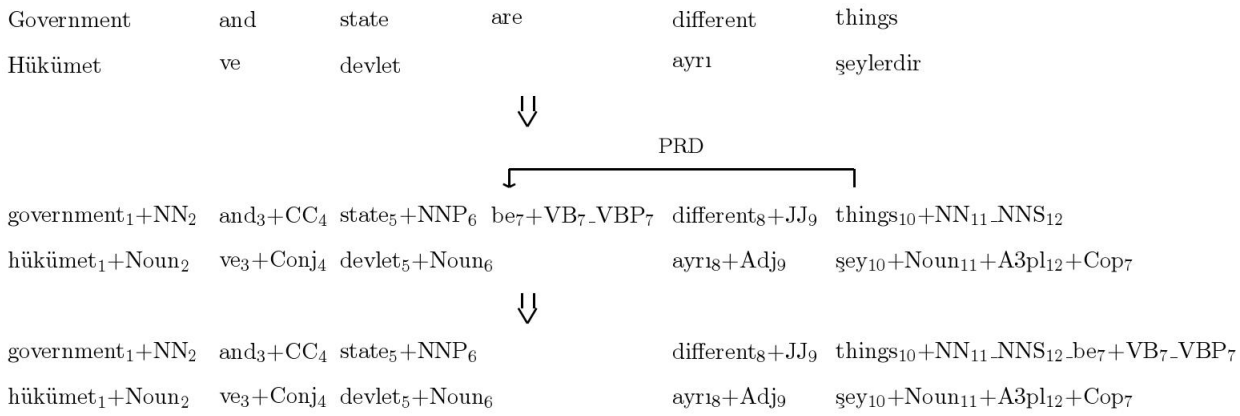


Figure 3.7: An example for copula transformation with predicate noun

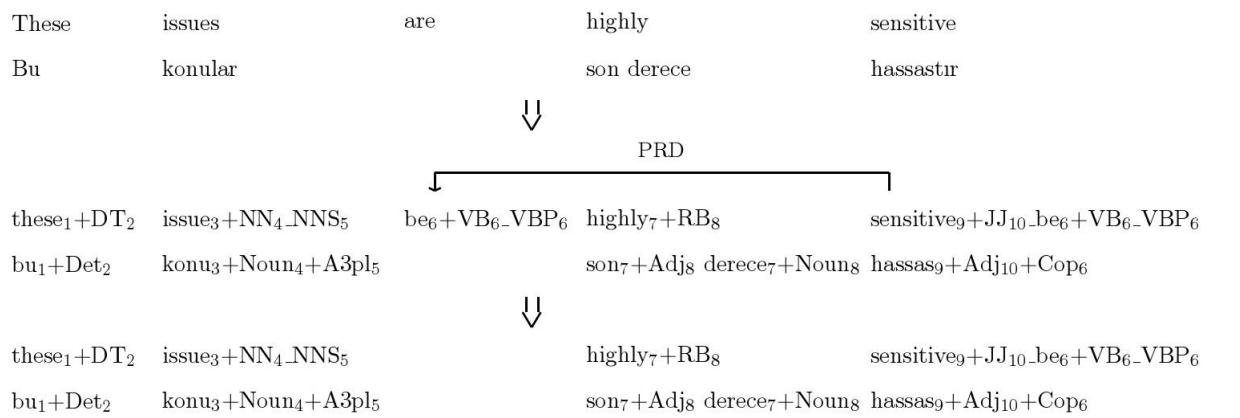


Figure 3.8: An example for copula transformation with predicate adjective

3.3.1.5 Auxiliary Verbs

An auxiliary verb is a function word which accompanies a verb. A lexical verb can take several auxiliary verbs which add different grammatical functions. In terms of dependency representation, each of these auxiliary verbs connects to the content verb with a VC (Verbal Chunk) label. In this section, we talk about each of these functions and related transformations.

Passive Voice

Passive voice is a syntactic transformation in which the subject is the target of the action that is denoted by the verb. In English, passive voice consists of an auxiliary verb (most of the time *be*) and the past participle form of the lexical verb. In Turkish the passive voice is represented with a *+Pass* morpheme in the verb. An example

transformation is given in Figure 3.9.

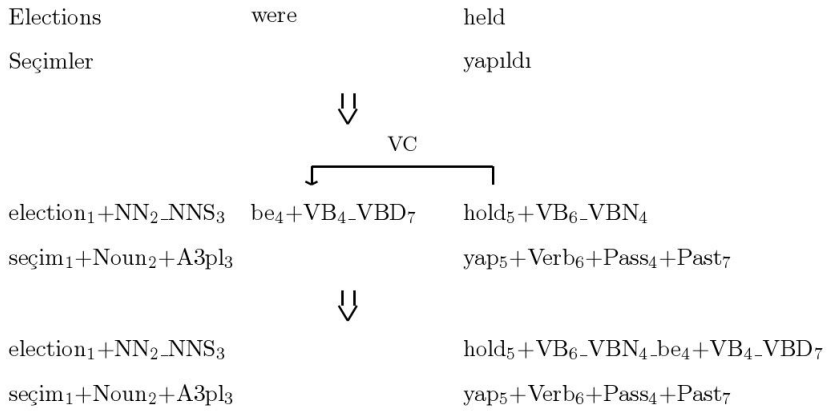


Figure 3.9: An example for passive voice transformation

Continuous Aspect

The continuous aspect is a grammatical aspect that expresses an ongoing occurrence of a state or event [Loos et al., 2003]. In English this is expressed with any conjugation of *be* together with the present participle form (ending with *-ing*) of the verb. In Turkish, this is mostly known as present continuous tense and mostly expressed with a suffix (*-(i)yor*) or any other *+Prog* morpheme (e.g.,*makta*). An example can be seen in Figure 3.10.

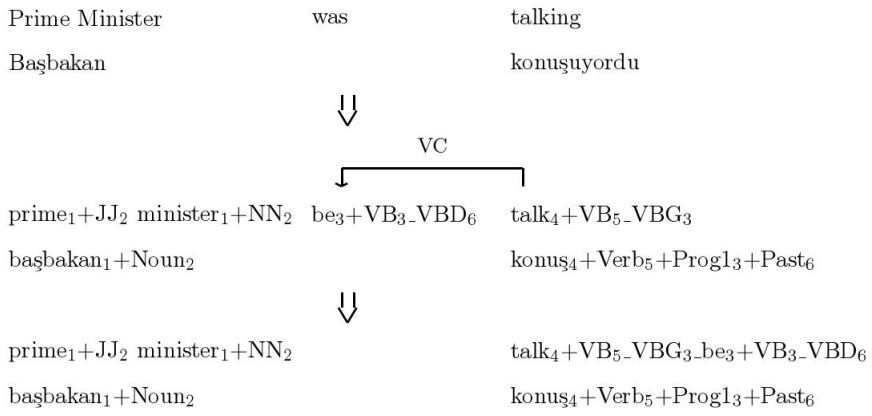


Figure 3.10: An example for continuous aspect transformation

Perfect Aspect

In perfect aspect, the focus is not just on the action of the verb, but also on the present

state arising from that action. In English, perfect aspect is formed by conjugating *have* and using it together with the past participle form of the verb. Similarly in Turkish, perfect aspect is usually formed by adding any *+Narr* morpheme to the verb. In our transformation we append *have* to the verb. Figure 3.11 gives an example to this transformation.

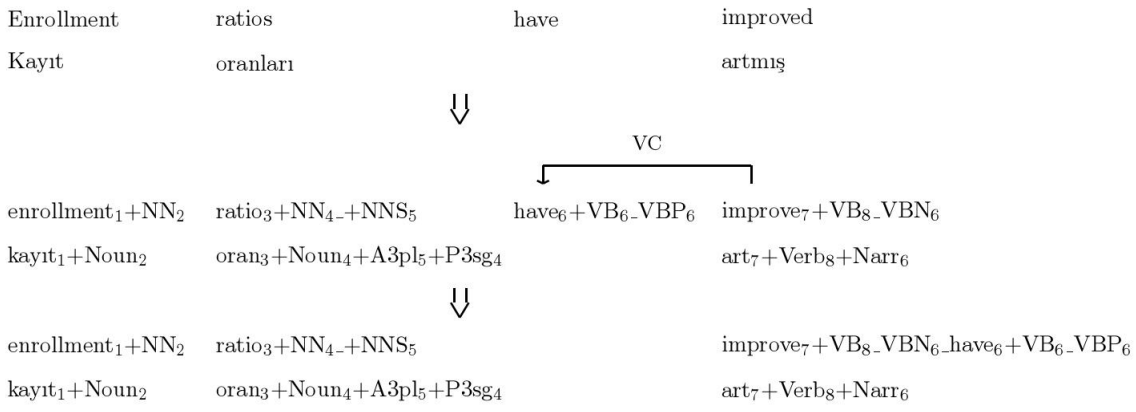


Figure 3.11: An example for perfect aspect transformation

Modals

A modal verb is a type of auxiliary verb which is used to indicate the modality of the verb. In English modals come before all the other auxiliary verbs and in Turkish they are represented with several morphemes. For instance, *will*, which is a commonly used modal, is used to indicate a future event and in Turkish *+Fut* morpheme is used to represent this. In the transformation step we append this modal to the main verb as seen in Figure 3.12.

Another widely used modal is the *can*. This is mostly used to express ability and in Turkish *+Able* morpheme is used for this purpose. Furthermore we use *must* to express an obligation or a necessity. In Turkish this same meaning is represented with *+Neces* morpheme. There are many other examples of such modals [Kerslake and Göksel, 2005].

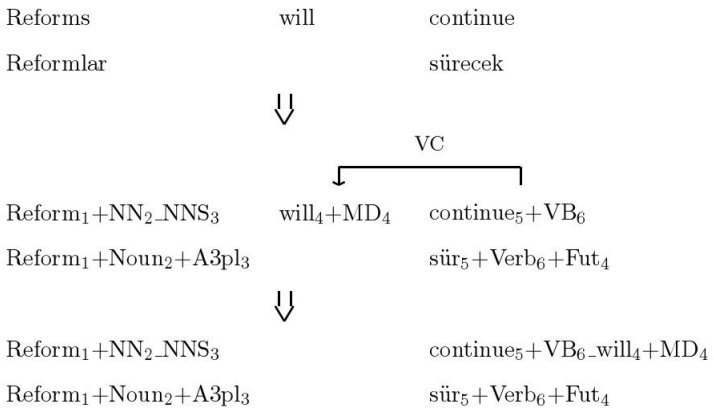


Figure 3.12: An example for modal transformation

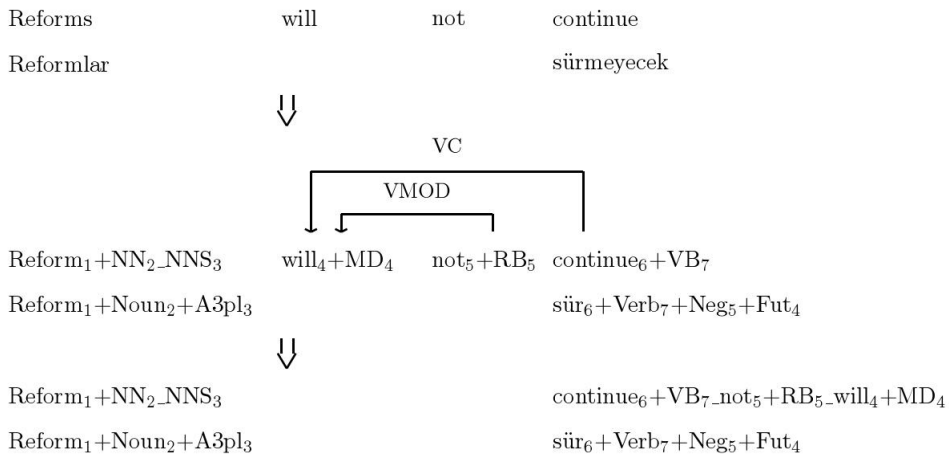


Figure 3.13: An example for negation transformation

3.3.1.6 Negations

Negation is a morphosyntactic operation which is used to invert the meaning of a lexical item [Loos et al., 2003]. In English, negation is performed with the negative particle *not* or its contracted form *n't*. In Turkish, a negative suffix is appended to a verb. An example transformation for negations is given in Figure 3.13.

3.3.1.7 Adverbial Clauses

An adverbial clause is a subordinate clause which functions as an adverb. It is a dependent clause so it cannot stand alone but is used together with another clause. It contains a subject and a predicate.

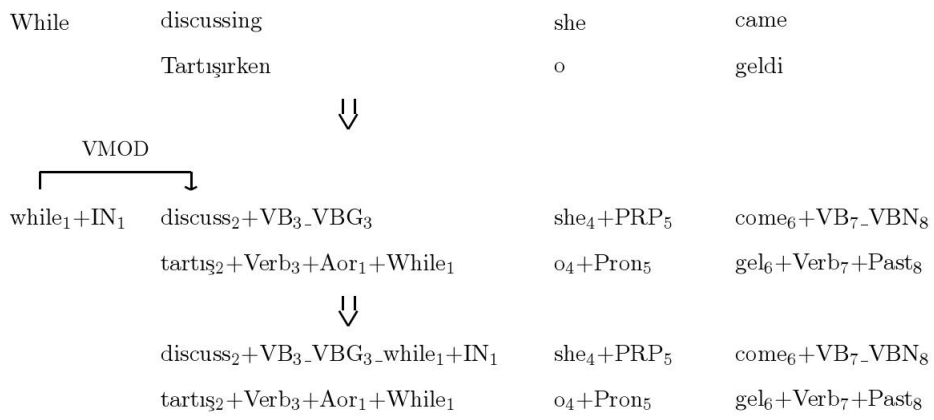


Figure 3.14: An example for adverbial clause transformation

In English, these clauses contain a subordinate conjunction which modifies the verbs. In Turkish, adverbial clauses take widely differing forms [Kerslake and Göksel, 2005]:

- Some clauses may be represented with a separate token without any morphological change such as

  [**As** there were going to be a lot of us,] I had bought another loaf.
  [Kalabalık olacağız **diye**] bir ekmek daha almıştım.

- Some clauses are translated into Turkish with a token and a morpheme appended to a verb:

  [**After** being repaired,] the machine broke down again.
  Makine [tamir edil**-dikten sonra**] yeniden bozuldu.

- Or some clauses are represented without any token but just morphologically

  Ahmet read that book [**when** he was a student].
  Ahmet o kitabı [öğrenci**-yken**] okudu.

We perform transformations on many of these cases. An example can be seen in Figure 3.14.

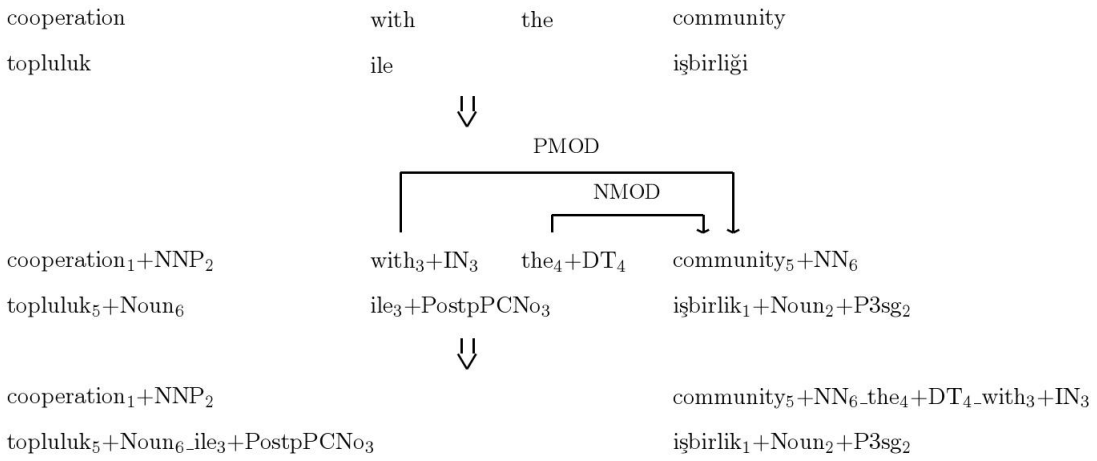


Figure 3.15: An example for postpositional phrase transformation

3.3.2 Turkish

3.3.2.1 Postpositional Phrases

In Turkish, although most of the grammatical relations are represented with morphemes, there are also a set of postpositions such as *ile (with)*, *için (for)*. Most of these postpositions correspond to the prepositions or subordinate conjunctions on the English side. Since we perform these preposition and subordinate transformations, we should make sure that the Turkish translations of these are in the same structure. In order to do this, we select the postpositions according to their frequency of usage of their English translations and append them to the related verb or noun like we did with English ones. Example transformations of postposition with a noun and a verb are given in Figures 3.15 and 3.16.

3.4 Experiments

We evaluated the effects of the transformations in factored phrase-based SMT with an English-Turkish data set which consists of 52712 parallel sentences. We partitioned this data into 3 sets; training set to generate the phrase-translation tables and generation tables, tuning set to optimize translation parameters and test set to evaluate the experiment. The tuning and test sets consist of randomly selected 1000 sentences. The

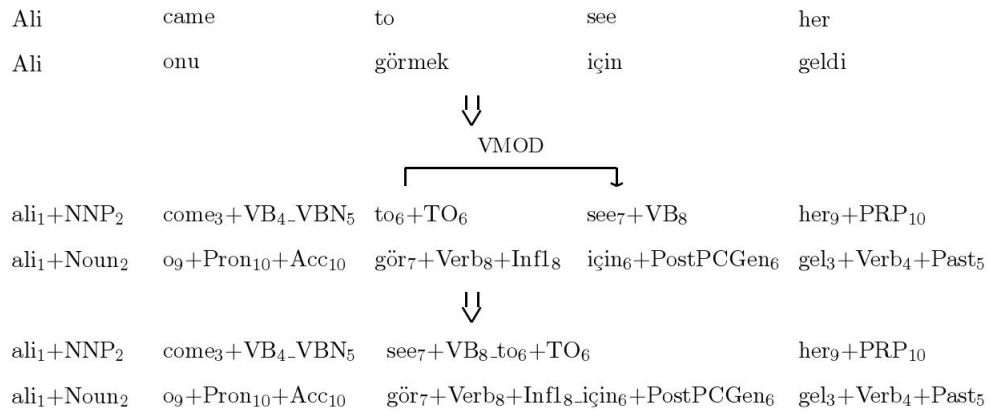


Figure 3.16: An example for postpositional phrase transformation

remaining of the sentences were used in the training.

To generalize the effects of the transformations we performed 10 trials for each experiment. We randomly generated these trial sets and used same sets in all of the following experiments.

We performed our experiments with the Moses toolkit [Koehn et al., 2007] which is a factored phrase-based beam-search decoder for machine translation. Moses is actually a complete SMT system which consists of all the necessary tools for training, decoding and evaluation. It uses the GIZA++ [Och and Ney, 2003], which is an implementation of the IBM Models, to establish the word alignments. From these word alignments Moses extracts the phrases. For our experiments, we limited the maximum phrase length to 7 which is the default value for Moses.

Furthermore, Moses works with any one of the three freely available language modeling toolkits which are SRILM [Stolcke, 2002], IRSTLM [Federico et al., 2008] and RandLM [Talbot and Osborne, 2007]. In this thesis we generated our language models with the SRILM toolkit. We produced 3-gram language models with Chen and Goodman's modified Kneser-Ney discounting (*-kndiscount* in SRILM) together with interpolation (*-interpolate* in SRILM).

In the decoding step, in order to allow for long distance reorderings we used a distortion limit[2] (*-dl* in Moses) of 40 and a distortion weight (*-weight-d* in Moses) of 0.1.

[2]Maximum number of words to skip in reordering

Finally for the evaluation of the results, we used the BLEU [Papineni et al., 2001] metric. For each experiment we gave statistics of BLEU scores such as maximum and minimum values, average and standard deviation.

3.4.1 The Baseline System

As a baseline system, we performed an experiment using the surface forms of the words without any transformation. In this experiment we used phrase-based approach with the 3-gram language model of surface forms. Table 3.3 shows the average, standard deviation, maximum and minimum BLEU scores for the 10 trials.

Experiment	**Ave.**	**STD**	**Max.**	**Min.**
Baseline	17.08	0.60	17.99	15.97

Table 3.3: BLEU scores for the Baseline System for 10 different train/test set

3.4.2 The Baseline-Factored System

We also tried our baseline system with a factored model. Therefore, instead of using just the surface form of the word, we put lemma, POS tag and morpheme information into the corpus. In factored translation, the factors are separated by a '|' symbol. Thus in this experiment we represented a token consisting of 3 factors as 'Surface|Lemma|POS_Morphemes'. An example to this representation is given in Table 3.4. In the baseline system, we used the first representation in Table 3.4 and in the baseline-factored system we used the last representation.

After preparing the data in above format, we aligned this parallel corpus based on the lemma factor because it is more general than the surface form. The rest of the factors were aligned accordingly. Furthermore, in factored models, user can generate different language models for different factors. We made use of this property and generated 3-gram LMs for each of the factors.

As Turkish is a morphologically rich language, we used a model that is similar to the one mentioned in Section 2.3.1. Instead of translating the surface forms, we

Representation	English/Turkish
Surface	relation+NN_NNS ilişki+Noun+A3pl
Lemma	relation ilişki
POS_Morphemes	NN_NNS Noun+A3pl
Surface|Lemma|POS_Morphemes	relation+NN_NNS|relation|NN_NNS ilişki+Noun+A3pl|ilişki|Noun+A3pl

Table 3.4: Several representations

translated lemma and POS_Morphemes separately and then generated the surface form. This approach is summarized in Figure 3.17.

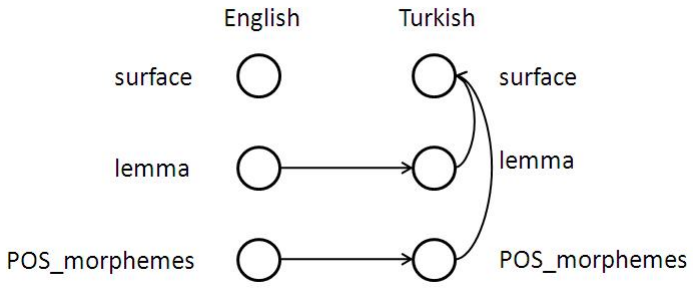


Figure 3.17: Translation by just using lemma and POS_morphemes

When we tried the model that is represented in the Figure 3.17, we got a score which is a little improvement to the baseline system as shown in Table 3.5. This is due to not using the already available information which is the translations of the surface form.

In order to prevent this information loss, we introduced an alternative path model which is illustrated in Figure 3.18. In this model, we first tried to translate the surface form. If we had a high probability surface form translation, we used it, otherwise we backed-off to lemma and POS_Morphemes information and generated surface form from the translations of those.

The results of this approach are given in Table 3.5. As you see, using lemma and POS_Morphemes information as a backup increases the results drastically. We continued using this alternative path model in the rest of the experiments.

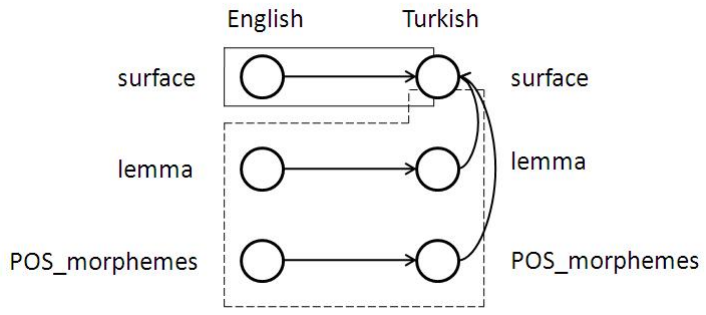


Figure 3.18: Alternative path model

Experiment	**Ave.**	**STD**	**Max.**	**Min.**
Baseline Model	17.08	0.60	17.99	15.97
Lemma and POS_Morphemes Model	17.55	0.65	18.46	16.26
Baseline Factored Model (alternative path model)	**18.61**	**0.76**	**19.41**	**16.80**

Table 3.5: BLEU scores of experiments with factored translation model

3.4.3 Noun-Adj

In order to see the effects of transformations separately, we performed them in several steps. In this first experiment, we only focused on the transformations that are performed on nouns and adjectives. When we performed these tranformations, our average BLEU score increased about 14% as seen in Table 3.6.

Experiment	**Ave.**	**STD**	**Max.**	**Min.**
Baseline-Factored	18.61	0.76	19.41	16.80
Noun-Adj	**21.33**	**0.62**	**22.27**	**20.05**

Table 3.6: BLEU scores for the baseline-factored and the noun-adj system

3.4.4 Verb-Adv

In next set of experiments, we focused on transformations that are performed on verbs and adverbs. Auxiliary verb and negation transformations are all performed on verbs and furthermore adverbial clause transformations are performed on adverbs. Table 3.7 contains the results of these experiments.

Experiment	Ave.	STD	Max.	Min.
Baseline-Factored	18.61	0.76	19.41	16.80
Verb	19.41	0.62	20.19	17.99
Adv	18.62	0.58	19.24	17.30
Verb+Adv	19.42	0.59	20.17	18.13
Noun+Adj+Verb+Adv	**21.67**	**0.72**	**22.66**	**20.38**

Table 3.7: BLEU scores for the verb-adv system with several combinations

From the above results, we can conclude that adverbial clause transformations (third row) are effective but not very consistent. Although change in the average score is very little, there may be some cases where the increase can be much larger, such as the 0.5 points improvement in the experiment with the minimum score.

The auxiliary verb and negation transformations improved the scores consistently which was expected due to the common and regular usage of auxiliary verbs. When we combined all these transformations (last row), we got the highest scores on average which is a 3.06 point improvement over the baseline-factored model.

3.4.5 Postposition (PostP)

Furthermore we also experimented with the postposition (PostP) transformations on the Turkish side. In Turkish, postpositions are mostly in adverbial clauses, therefore to see the relationship between postposition transformations in Turkish and adverbial clause transformations in English, we performed several experiments which include and exclude these transformations. Table 3.8 summarizes the results of these experiments.

Experiment	Ave.	STD	Max.	Min.
Noun+Adj+Verb	21.75	0.71	23.07	20.70
Noun+Adj+Verb+PostP	21.89	0.66	22.88	20.66
Noun+Adj+Verb+Adv	21.67	0.72	22.66	20.38
Noun+Adj+Verb+Adv+PostP	**21.96**	**0.72**	**22.91**	**20.67**

Table 3.8: BLEU scores of postposition experiments

In Table 3.8, the first two rows are the cases in which the adverbial (Adv) transformations were excluded. In this case we saw that postposition transformations improve the score 0.14 points from 21.75 to 21.89. On the other hand, the last two rows show

the experiments when the Adv transformations were included. According to these two experiments using postposition transformations made an increase of 0.29 on average, which is more than twice the increase we got before. Therefore we can conclude that the adverbial clause transformations and the postposition transformations have a positive effect on each other.

3.5 Discussion

In order to see the relative improvement of each experiment, we drew the graph in Figure 3.19. In this graph the experiments are ordered according to their average scores. For each experiment the average, maximum and minimum BLEU scores over 10 experiment are represented in the graph.

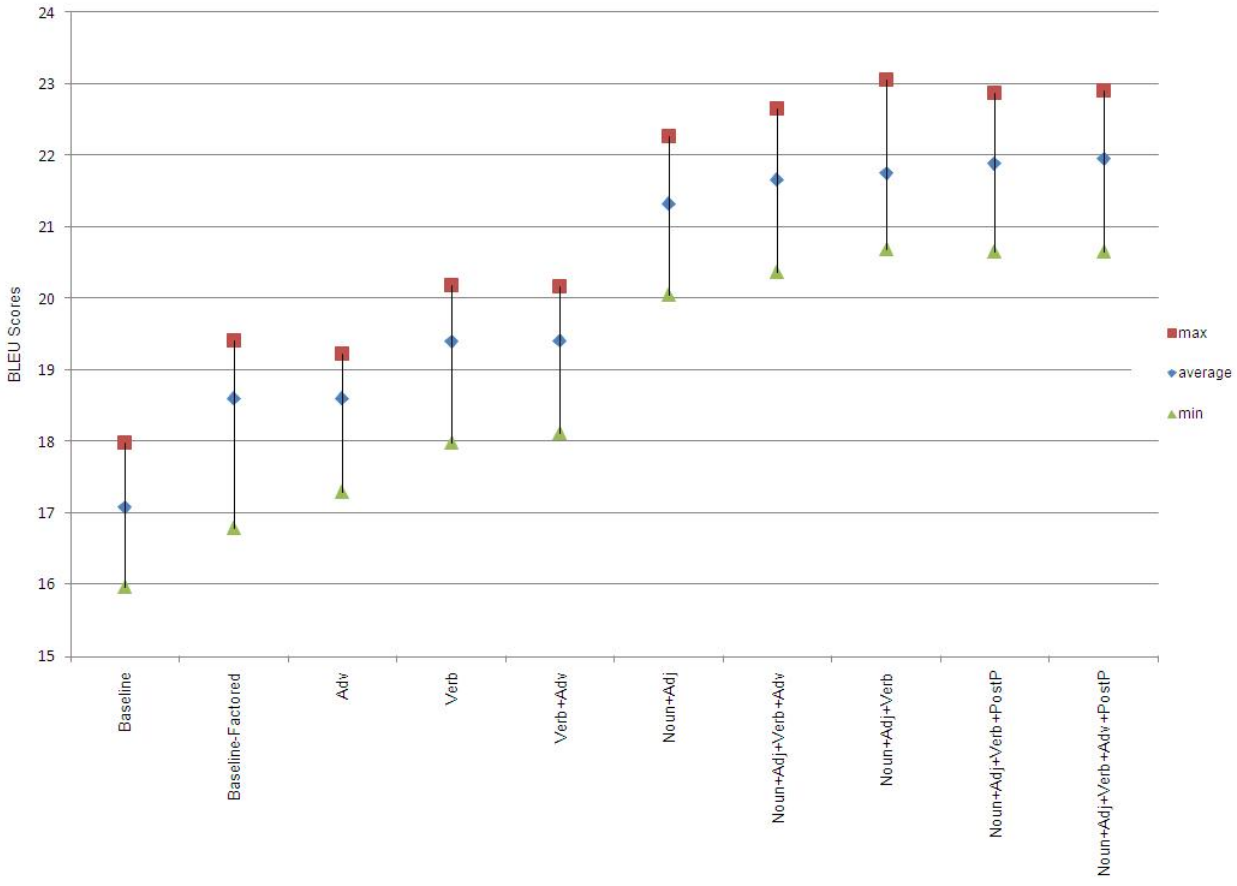


Figure 3.19: BLEU scores of each experiment

Moreover, Figure 3.20 represents the change in BLEU scores of all 10 experiments for each case. In this graph, we see that the change in BLEU scores is mostly consistent for different train/test set partitioning. Therefore, we can continue our discussion according to Figure 3.19.

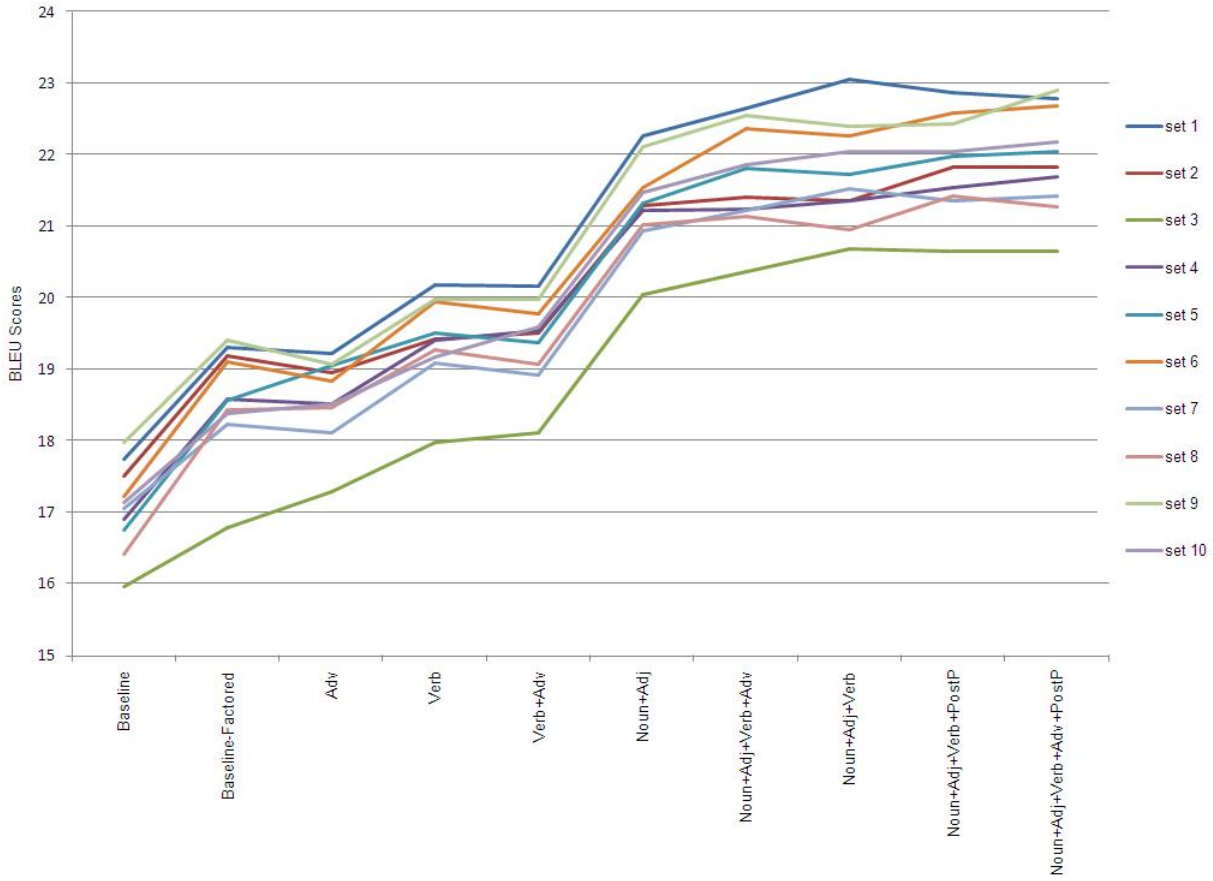


Figure 3.20: BLEU scores of 10 experiments for each case

We started our experiments with a baseline of 17.08 BLEU points. We got an improvement of 1.53 points when we started using the factored model with alternative paths. This improvement is most likely due to the two important advantages of factored translation model: The first one is the back-off mechanism of translating lemmas in case a good surface translation is not available. In addition to this generalization gain, factored models also help in reordering. By using different language models with lemma and especially with POS_Morpheme factor, we are able to include more syntactic features in our reordering. This is another benefit of factored translation models.

The rest of the experiments in Figure 3.19 are all about the transformations. We see that every transformation improves the existing system and the highest performance is reached when all transformations are performed. However when we take a closer look at the individual transformations performed on English side, we observe that not all of them have the same effect. While Noun+Adj transformations give us an increase of 2.73 BLEU points, transformations of Verbs improve the result by only 0.8 points and improvement with adverbs is even lower. In order to understand why we get such a difference, we looked at the number of tokens.

Table 3.9 contains some word statistics of our data before any transformation is performed. We can note that Turkish has twice the number of distinct words than English, but Turkish has smaller number of distinct lemmas. This difference is due to the rich morphological structure of Turkish.

	Sentences	**Words**	**Unique Words**	**Unique Lemmas**
English	52,712	1,205,347	29,232	18,282
Turkish	52,712	942,420	60,452	16,771

Table 3.9: Statistics on English and Turkish data

Similarly the use of function words in English causes a big difference in the number of words between Turkish and English, as shown in Table 3.9. During the transformations, we appended these function words to the related content words, so the number of tokens decreased as we perform these transformations. The graph in Figure 3.21 illustrates the BLEU scores and the remaining number of tokens as we performed our transformations. From the graph we can see that as the number of tokens in English decrease, the BLEU score increases. In order to measure the relationship between these two variables statistically, we perform a correlation analysis and find that there is a strong negative correlation of -0.99 between the BLEU score and the number of tokens. This means that, as we append the function words to the related content words and hence reduce number of words in English, the languages are structurally getting closer and as a result, we get better translations. Therefore syntax to morphology alignment works as intended.

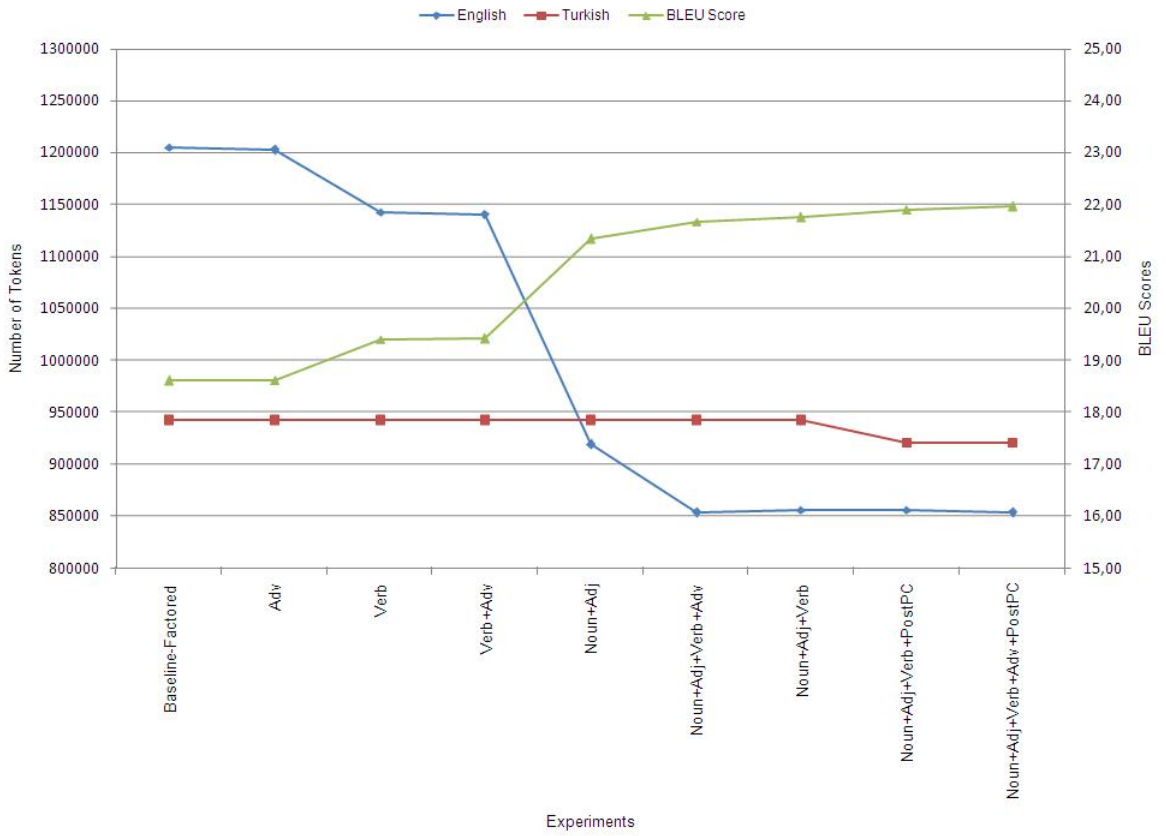


Figure 3.21: Relation of BLEU scores with number of tokens

3.6 Related Work

Statistical Machine Translation into a morphologically rich language is a challenging problem because current SMT systems cannot perfectly deal with the complex morphology of these languages. A good example for this is given by Koehn for Finnish which is also a morphologically complex language [Koehn, 2005]. In this paper, the author collected parallel texts in 11 languages from the proceedings of the European Parliament. Phrase-based SMT systems were developed between these pairs and among the 10 SMT systems which uses English as a source language, Finnish returns the worst BLEU score, which was 13.00 while the average score was 23.84. This proves that standard methods are not enough while working with morphologically rich languages.

In order to improve the translation quality of these systems, several approaches have been tried. One significant work in this area was performed by Oflazer and El-Kahlout [Oflazer and El-Kahlout, 2007]. In this work, the authors explored different

representational units of the lexical morphemes. They showed that the segmented Turkish morphemes are aligned with the function words of English despite the limited data they have. With some additional improvements, the authors were able to improve their system from 20.22 BLEU points to 25.08 BLEU points. Our approach is different from their work. They mostly work on the Turkish side; however, we approach the problem from reverse direction and perform most of our transformations on the English side. They experimented with different segmentations of Turkish morphology, alternatively we work on appending function words to their related content words in English.

Factored translation models have been used in several SMT systems. One early work was published by Koehn and Hoang [Koehn and Hoang, 2007]. In this paper, the authors used the alternative path approach we used, on their German-English data. When they used the lemma/morphology model to generate surface forms as a backoff to the surface translation model, they improved their score from 18.19 BLEU points to 19.47 BLEU points.

Moreover, similar to us, Birch *et al.* integrated more syntax into the factored translation model by using Combinatorial Categorial Grammar (CCG) supertags as a separate factor [Birch et al., 2007]. These tags provide a rich source of syntactic information by containing the syntactic context of words. Using this factor together with the POS and word factors resulted in an increase of 0.46 BLEU points from 23.97 to 24.43. Our approach is different from these models in terms of our third factor. As one can remember we use actual dependency relations in order to produce our POS_Morphemes factor of the source side. The idea of using dependency in syntax to morphology alignment has been first applied in this thesis.

Chapter 4

SYNTACTIC REORDERING

4.1 Motivation

Handling word order differences between languages is still one of the remaining challenges in MT. Even the state-of-the-art MT systems cannot completely solve this problem. Phrase-based translation models can only capture some common local reorderings within phrase pairs, but for long-range reorderings, these systems cannot do much. The reason for this is that phrase-based systems cannot use syntactic information, which seems to be very crucial while dealing with reordering. In recent years, many MT researchers have worked on methods to incorporate this syntactic information into MT systems [Collins et al., 2005, Xu et al., 2009].

Reordering problem is important because many language pairs, even the close ones, have word order differences. Distant languages such as English and Turkish have different word order patterns. While Turkish has a flexible SOV constituent order, English is rather fixed on using SVO. Additionally, Turkish adverbs mostly precede the verb, but in English they come after the verb and most of the time even after the object. Moreover, some other ordering differences occur while using the subordinate clauses and passive voices. In order to deal with these differences, we propose an approach which incorporates syntactic information into a phrase-based SMT system.

4.1.1 Overview of the Approach

Our syntactic reordering approach is similar to our syntactic transformations. We perform our reorderings in the source side so that the word order becomes similar to its target sentence in the corpus. After performing the syntactic transformations that we mention in Chapter 3, we execute our reordering methods on English sentences then give them to the SMT system.

Our approach starts with grouping the consecutive words according to their dependencies. By using the deprel tag of the dependency parser output, we label these groups whether they are subject, object, verb, adverb or subordinate clause. Finally we use some rules to relocate these groups. An example reordering is given in next section.

4.1.2 An Example

Let's assume we are given the below sentence pair.

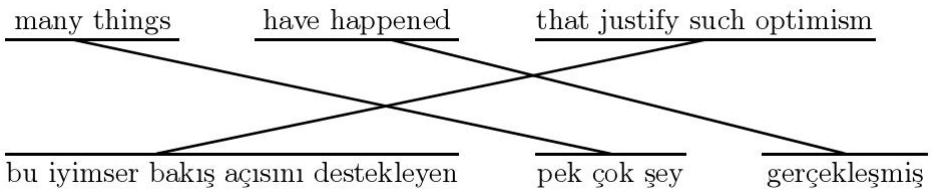
Many things have happened that justify such optimism
Bu iyimser bakış açısını destekleyen pek çok şey gerçekleşmiş

Before starting our syntactic reordering, we will perform the necessary analysis and syntactic transformations on these sentences. In order to give the reader a simple and clear representation, we will continue our example over this untouched surface form as opposed to the morphemic representation used earlier.

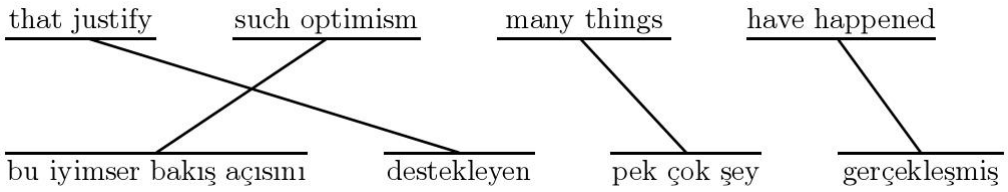
Our first step is to group the words in English side. At the end of this step we will get the three groups below.

Many things have happened that justify such optimism

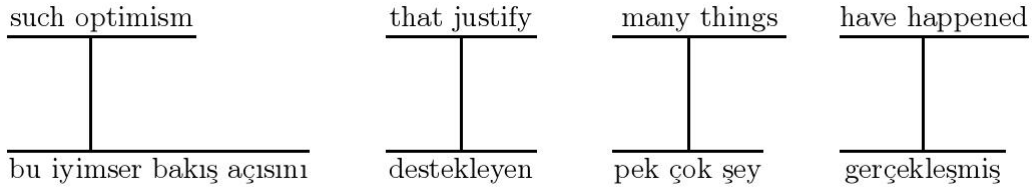
In the sentence above, the first group of words form the subject and the second group forms the verb complex. As you see there is the subordinate clause "that justify such optimism" which is attached to the main clause. The determiner *that* stands for the subject of our sentence, so this clause modifies the subject. If we align these groups with their Turkish correspondences, we will get something like



In the above alignment, the link between the aligned subordinate clauses cause a crossing with the other links. We know that in Turkish, a noun modifier comes just before the noun phrase. Therefore, if we put the subordinate clause just before the subject, we will get the same order with the Turkish translation. With this movement, we perform the necessary top level reordering. However, there is still some work that needs to be done in the subordinate clause. If we group the words in the subordinate clause, our alignment will become



Here "such optimism" is the object of this clause. If we move this object before the verb, we will again get a *monotonic* alignment with the Turkish sentence.



As a result, we perform two reorderings in this sentence to make it structurally similar to its Turkish translation. Since our algorithm works recursively, there may be more than one reordering in a sentence.

4.2 Reordering Constituents

In Turkish, variations in word order together with the position of the stress affect the meaning of the sentence [Kerslake and Göksel, 2005]. Although there are many word order variations, in this thesis we develop our rules based on the most common patterns. In this section, we describe these reorderings that we perform on English sentences in detail.

4.2.1 Object Reordering

The most commonly used difference between English and Turkish sentences is the ordering of the object. In English, the object comes after the verb (SVO) but in Turkish most of the time the object is placed between the subject and the verb (SOV). In order to have a similar word ordering, we move the whole phrase of the object to the place just before the verb. This reordering is performed in main clause and subordinate clauses. An example is given in Figure 4.1.

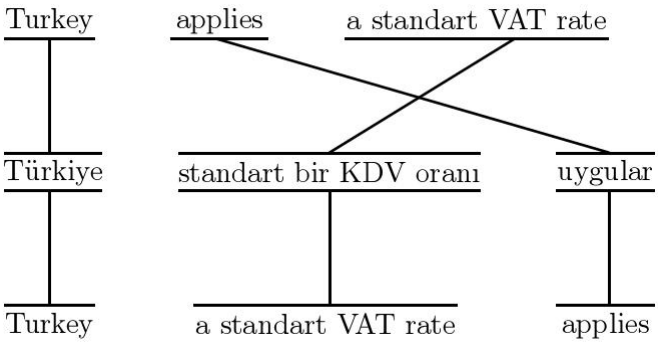


Figure 4.1: An example for object reordering

In Figure 4.1, the first line is the original English sentence without any reordering. Here the labels of the groups are in order from left to right: subject, verb and object. The second line is the Turkish translation of the first line. As it can be observed from the alignments, the order is in SOV form. Because of this ordering difference we observe crossing alignments. If we reorder the object in the English sentence, we will get the last line. Since the alignments are monotonic, there are no crossing lines in the alignment.

4.2.2 Adverb Reordering

There are several adverb types such as adverbs of manner, adverbs of place, adverbs of time, etc. In a language, the locations of adverbs may vary according to these types. Furthermore, the locations of these adverbs change from language to language. In English, most of these adverbs are placed after the object or after the verb if there is no object. In Turkish, the locations of adverbs vary a lot around the subject therefore we move English objects to right behind the subject. An example adverbial reordering is given in Figure 4.2.

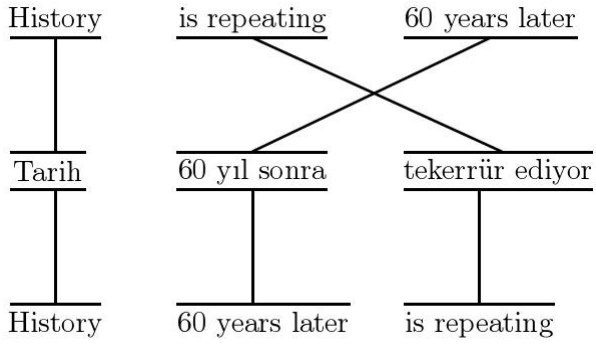


Figure 4.2: An example for adverb reordering

4.2.3 Passive Voice Reordering

Passive voice is used in order to emphasize the receiver and suppress the agent of the action. However, the agent can still be expressed in the sentence. In English, this agent is expressed after the verb, but in Turkish, it comes immediately before the verb. In order to have a similar word order, we also put these agents just before the verb in English sentences. An example to this reordering can be seen in Figure 4.3.

4.2.4 Subordinate Clause Reordering

In English, subordinate clauses that have the conjunctions *that* or *which* are used in order to modify the noun they are referring. In Turkish, a noun modifier comes before the noun. Therefore, we first find the noun phrase that is modified and then put the

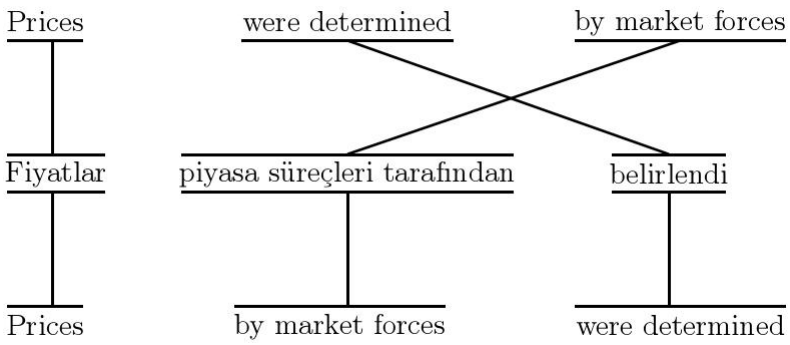


Figure 4.3: An example for passive reordering

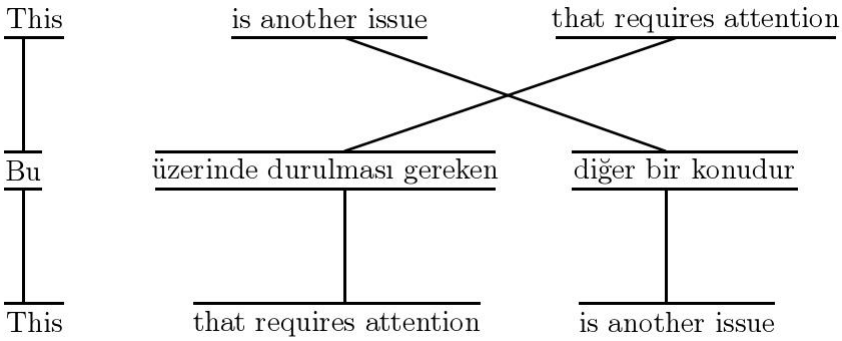


Figure 4.4: An example for subordinate reordering

whole subordinate clause before this noun. An example to this reordering is given in Figure 4.4.

4.3 Experiments

We performed reorderings with the same 10 sets we used in the transformation experiments. Furthermore, we continued using the same tools with the same parameters.

As our baseline system we used the best transformation output we got from the previous chapter, which is the Noun+Adj+Verb+Adv+PostPC output. We started with the top level object reordering. Results of this experiment are summarized in Table 4.1.

As one can see from the table, the scores are really close. Five times out of the ten trials, the object reordering got higher scores than the system with no reordering. Since the scores are very close, we continued using object reordering in the remaining

Experiment	Ave.	STD	Max.	Min.
No Reordering (Noun+Adj+Verb+Adv+PostPC)	21.96	0.72	22.91	20.67
Obj	21.94	0.71	23.12	20.56

Table 4.1: BLEU score of the object reordering experiment

of the experiments to see if object reordering will become more effective with the other reorderings. Table 4.2 shows combinations of object reordering with the rest of the reorderings.

Experiment	Ave.	STD	Max.	Min.
Obj	21.94	0.71	23.12	20.56
Obj+Adv	21.73	0.50	22.44	20.69
Obj+Passive	21.88	0.73	23.03	20.51
Obj+Subord	21.88	0.61	22.77	20.92

Table 4.2: BLEU scores of all experiments

As represented in Table 4.2 we see no improvements with the other reorderings. There may be a couple of reason for this. We discuss these possible reasons in detail in the next section.

4.4 Discussion

There may be possible explanations for the results in Table 4.2. One explanation for the adverbs can be the variations on the Turkish side. In Turkish, adverbs may be used in different places according to the emphasis of the sentence. Our adverb reordering may not be covering all these variations.

Another reason why the reorderings are not statistically significant, can be the frequency of their usage. Table 4.3 summarizes the frequencies of these reorderings and unfortunately, according to this table, there seems to be no relation between these frequencies and the BLEU scores.

Reorderings do not seem to improve our test scores but we wonder if they improve the alignment quality by producing more monotonic alignments. Therefore, we looked

	Frequency
Obj	15,804
Adv	5,078
Passive	2,778
Subord	3,481

Table 4.3: Numbers of time different reorderings are applied

at the alignment files that were created by GIZA during the training. We used two metrics on these alignments. Our first metric was the *absolute distance* metric which finds the absolute distance between the positions of two tokens of an alignment. Our other metric was the *crossing alignments* metric. In this metric, we counted the number of times the links of alignments cross with each other. In case of a monotonic alignment both of these metrics will return a number close to zero since the positions of aligned words in sentences will be close to each other with no other alignment link crossing the other alignment's link.

We used these two metrics on our experiment files and got the results shown in Table 4.4. This table tells us that on the average an alignment link can be crossed with 3 or 4 other alignment links. Furthermore, distance information indicates that on the average the alignment of the i_{th} word is most probably somewhere close to $(i+6)_{th}$ or $(i-6)_{th}$ position of the translated sentence.

Experiment	**Crossings**	**Distance**
No Reordering	3.45	5.56
Obj	3.40	5.54
Obj+Adv	3.40	5.54
Obj+Passive	3.39	5.54
Obj+Subord	3.37	5.50

Table 4.4: Average number of crossings and average absolute distance

As shown in Table 4.4, all reorderings are slightly reducing both metrics. Due to the high frequency of objects, the reduction for object reordering is more than the others. Similarly the change with subordinate clause reordering is more than the change with adverb or passive reordering. The reason for this may be the length differences of these phrases. In subordinate clause, we move a whole clause while in adverb or

passive, reordering is just limited with a couple of words.

As you can remember, we have performed all of our reordering experiments on the Noun+Adj+Verb+Adv+PostP model. The reason why we did not get much improvement with reordering may be because we may have already performed the necessary improvements with transformations. Therefore, in order to see whether transformations have any effect on reordering, we also performed reorderings on the first 2 sets of the baseline-factored model. The results of these experiments are summarized in Table 4.5.

	No Reord.	**Object Reord.**	**Passive Reord.**	**Subordinate Reord.**
Set 1	19.32	19.63	18.98	19.32
Set 2	19.20	18.93	18.97	18.62

Table 4.5: Average BLEU scores for reorderings on baseline model

According to the table, reorderings do not produce a consistent change in the results therefore we can conclude that transformations do not have any effects on reorderings.

As a result, we observe that the syntactic reordering did not make a statistically significant improvement with our limited training data. These reorderings may become more dominant in phrase tables when more training data is available. Unfortunately, there is not much prior result for reordering on English-Turkish pair. One recent work [Xu et al., 2009], was able to get an improvement of 0.6 BLEU points by using preprocessing reordering as we did. However we should note that, in that experiment the authors work with a parallel data that consists of 76M words while our corpus have just 1,205,347 words in English.

4.5 The Contribution of LM to Reordering

Lastly, we investigated the contribution of using a higher order n-gram LM. As one can remember, we can specify different LMs for different factors. Similarly we used a 3-gram LM for the POS_Morphemes factor. In this experiment, we investigated whether using a high order LM with this factor will improve the reordering by including longer

syntactic features to the system.

In this experiment, we continued using Noun+Adj+Verb+Adv+PostP model which is our current best model. Moreover, training set and same parameters were continued to be used to train the LMs. We searched for the optimum n-gram order over 2 sets instead of 10. Figure 4.5 represents the BLEU scores of these experiments with different n-gram orders.

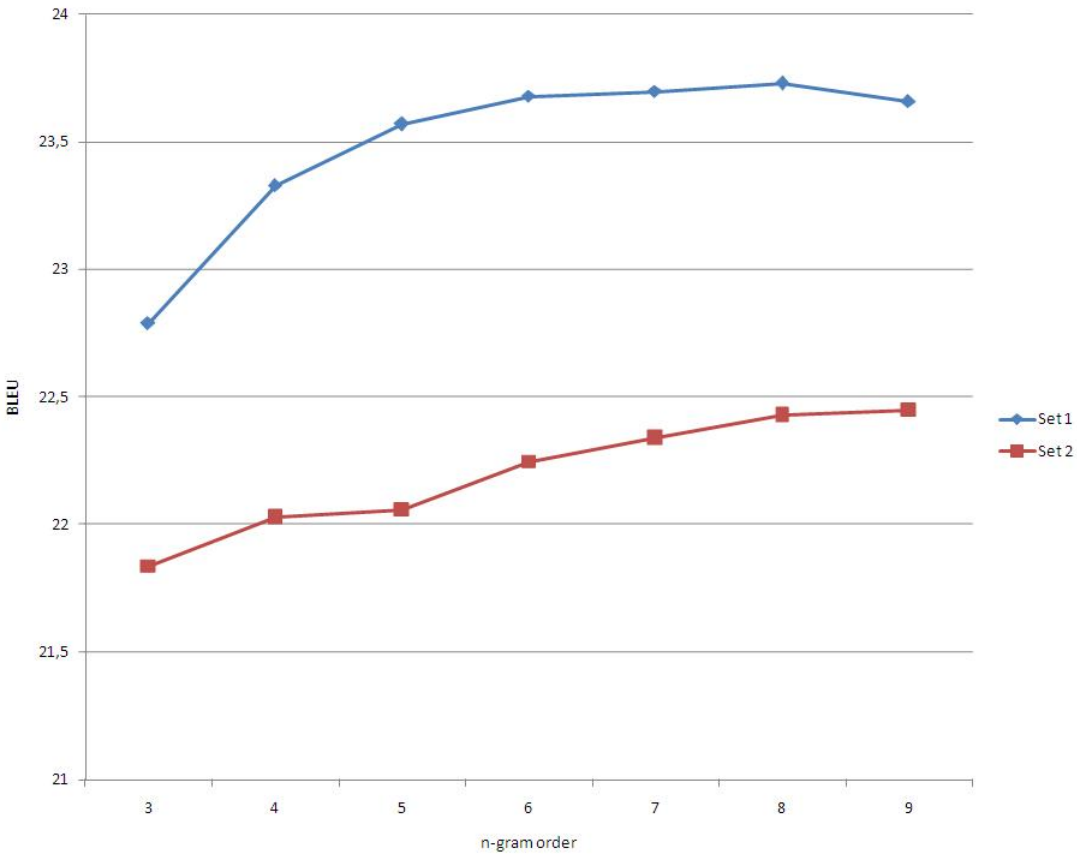


Figure 4.5: BLEU Scores with different n-gram orders

In Figure 4.5, as we increase the order of the LM, we also observe an improvement in the BLEU score. Of course this increase cannot continue forever. We saw our first decrease in BLEU at order 9 therefore we stop there and decided to use 8-gram LMs. When we applied 8-gram LM for POS_Morphemes factor with Noun+Adj+Verb+Adv+PostPC model, the average BLEU score increased 0.65 points and became 22.61 as seen in Table 4.6.

Experiment	Ave.	STD	Max.	Min.
3-gram	21.96	0.72	22.91	20.67
8-gram	22.61	0.77	23.66	21.37

Table 4.6: BLEU score for different order LMs

4.6 Augmenting the Training Data

In order to alleviate the lack of parallel corpora, we performed some augmentations on training data. First of all, we added some commonly used phrases and words, such as day and month names or numbers, to the training corpus. These additions did not produce a significant improvement. Therefore we augmented the training data with reliable phrase-pairs that are obtained from the earlier phrase table. We chose these phrase-pairs according to their phrase translation probabilities. We assume that a phrase-pair is reliable if the ratio of the probabilities is larger than 0.9 and smaller than 1.1, and the sum of these probabilities is larger than 1.5. Phrase-pairs which satisfy these conditions are added to the training corpus. The effects of these augmentations are summarized in Table 4.7.

Experiment	Ave.	STD	Max.	Min.
without any augmentation	22.61	0.77	23.66	21.37
with some common phrase-pairs	22.60	0.76	23.89	21.00
with reliable phrase-pairs	23.78	0.71	24.52	22.25

Table 4.7: BLEU score of the experiments with the augmented training data

As it is observed from the table, addition of the reliable phrases to the training corpus improves the performance significantly. Adding these phrase-pairs will improve the probability of them and hence result in the selection of these phrases more.

4.7 Some Sample Translations

Until now we have reported the effects of our methods in terms of BLEU score. In order to see the improvements in translation quality, we also looked at the produced translations. In this section we are going to give two sample translations.

Input: `The relevant conventions of the Council of Europe have not yet been ratified.`
Reference: `Avrupa konseyinin ilgili sözleşmesi henüz onaylatılmamıştır.`
Translation of baseline system: `Avrupa konseyi sözleşmesinin ilgili henüz onaylatılmamıştır.`
Translation of our best system: `Avrupa konseyinin ilgili sözleşmesi henüz onaylatılmamıştır.`

Input: `Administrative capacity in different areas need to be strengthened to ensure that the acquis is implemented and enforced effectively.`
Reference: `Müktesebatın etkili biçimde uygulanmasını temin etmek üzere, değişik alanlardaki idari kapasitenin güçlendirilmesi gerekmektedir.`
Translation of baseline system: `Idari kapasite değişik alanlarda, müktesebatın yürütülmesi ve etkili bir biçimde uygulanması temin güçlendirilmesi gerekmektedir.`
Translation of our best system: `Müktesebatın etkili biçimde uygulanmasını temin etmek üzere, değişik alanlardaki idari kapasitenin güçlendirilmesi gerekmektedir.`

In above examples we can see the input English sentence, its reference Turkish translation which was produced by a human translator, a Turkish translation of the baseline system and another Turkish translation produced by the improved system. In both of these examples our improved system was able to produce the same exact translation that is given by the human translator while the baseline system cannot.

In the first example, the words in the sentence that is produced by the baseline system are correct but the morphology and the word ordering is wrong. In this example, we see that our transformation approach work as intended and produce the right surface form for the words `konseyinin` and `sözleşmesi`. Furthermore using a higher order LM may be the reason why we get the correct word ordering. The second example has also the similar problems which were corrected with our methods. More detailed representations of these sentences are given in Appendix A.

4.8 Related Work

Several approaches have been tried to integrate syntactic information into SMT systems in order to improve reordering. One method which is similar to our approach was the work of Collins *et al.* [Collins et al., 2005]. In this paper, the authors attempted to reorder the constituents of source language in such a way that the word order is very similar to the corresponding sentence in target language. They performed reordering by applying a series of manually crafted rules to the syntactic parse tree of source sentence. They tried their approach on translation from German to English over 750K sentence pairs and increased their BLEU score from 25.2 to 26.8. Although, this paper is similar to our work in terms of applying source sentence reordering using syntax. They used syntactic parse tree rather than dependency parser.

Another work which applied a preprocessing reordering is the work of Xu *et al.* [Xu et al., 2009]. In this recent paper, the authors presented a precedence reordering approach based on a dependency parser. Manually extracted precedence rules were used during the reordering. They applied their approach in translation from English, which is a SVO language, to five SOV languages which include Turkish. In English to Turkish experiment, their reordering approach improved their baseline from 9.8 BLEU point to 10.4 with a 76M words of parallel data. This work is superficially similar to ours in terms of using a dependency parser. The authors performed similar top level reordering as we did.

Moreover, there are some reordering approaches which integrate syntactic information into SMT system through factored translation models. For instance, Hoang and Koehn used POS tag translations in factored translation models to improve mid-range reorderings [Hoang and Koehn, 2009]. During the decoding step, they used the extra POS tag translation model to create templates for surface word translations. With this approach the authors obtained a 1.0% BLEU increase. This method is different than ours in terms of using POS-based reordering instead of dependency-based reordering. Furthermore, this statistical approach does not require any reordering rules like ours. On the other hand using this approach for Turkish is a bit challenging due to the complex morphological structure. In Turkish a word can be a noun or an adjective at the

same time depending on the surrounding words.

Chapter 5

SUMMARY AND CONCLUSIONS

In this thesis, we have introduced a novel approach to align English syntax with Turkish morphology via using dependency relations on the English side so that the structure of English sentences become more close to Turkish structure. We have applied transformation rules in order to associate function words to their related content words. With these transformations we were able to capture the relationship between English syntax and complex morphological tags of Turkish words and improved the translation quality by 3.35 BLEU points. Furthermore, we showed that as we performed the transformations and reduced the number of tokens in English, the BLEU score increased equally. As a result, these findings strongly support this new alignment approach between syntax and morphology.

Moreover, we have performed syntactic reorderings in source side in order to make its word order closer to the word order of the target language. We have identified and reordered the constituents of the source sentence by using the dependency relations. Although we have seen some improvements in terms of alignments, the reorderings did not produce a statistically significant improvement in translation quality. In order to see the effects of reordering more clearly, we need to apply these to a significantly larger corpus.

Since we do not expect that kind of large parallel corpus in near future, we tried to improve ordering with another approach. We made use of one of the advantages of factored translation models which is specifying different LMs for different factors. We used larger n-gram LMs with POS_Morphemes factor in order to incorporate the

available syntactic information in this factor more into the translation system. This approach produced some significant improvements in our translations and resulted in an increase of 0.65 BLEU points.

Another advantage of factored translation models was the generalized back-off model which is translating lemma and POS_Morphemes separately and then generating the surface form. With this approach, we were able to translate most of the words that are not seen in training but exist among the test sentences and increased our performance 1.06 BLEU points.

The proposed approaches in this thesis were tried on SMT from English to Turkish and improved the translation quality by overall 6.7 BLEU points. As the future work we are going to apply these methods to Finnish which is another language with complex morphological features and to Arabic.

Chapter A

APPENDIX A

A.1 Example 1

English sentence in its surface form:

`The relevant conventions of the Council of Europe have not yet been ratified.`

English sentence after the POS tagging:

`the+DT relevant+JJ convention+NN_NNS of+IN the+DT Council+NNP of+IN Europe+NNP have+VB_VBP not+RB yet+RB be+VB_VBN ratify+VB_VBN .+.`

English sentence after the dependency parsing:

```
1   the          the         DT   DT   _  3   NMOD
2   relevant     relevant    JJ   JJ   _  3   NMOD
3   conventions  convention  NNS  NNS  _  9   SBJ
4   of           of          IN   IN   _  3   NMOD
5   the          the         DT   DT   _  6   NMOD
6   Council      council     NNP  NNP  _  4   PMOD
7   of           of          IN   IN   _  6   NMOD
8   Europe       europe      NNP  NNP  _  7   PMOD
9   have         have        VBP  VBP  _  0   ROOT
10  not          not         RB   RB   _  9   VMOD
11  yet          yet         RB   RB   _  9   ADV
12  been         be          VBN  VBN  _  9   VC
13  ratified     ratify      VBN  VBN  _  12  VC
14  .            .           .    .    _  9   P
```

English sentence after transformations:

`relevant+JJ convention+NN_NNS_the+DT council+NNP_of+IN_the+DT europe+NNP_of+IN yet+RB ratify+VB_VBN_have+VB_VBP_not+RB_be+VB_VBN .+.`

Translated Turkish sentence in its morphological form:

```
Avrupa+Noun konsey+Noun+P3sg+Gen ilgili+Noun sözleşme+Noun+P3sg henüz+Adverb
onayla+Verb+Pass+Neg+Narr+Cop .+Punc
```

Translated Turkish sentence in its surface form:

```
Avrupa konseyinin ilgili sözleşmesi henüz onaylatılmamıştır.
```

A.2 Example 2

English sentence in its surface form:

```
Administrative capacity in different areas need to be strengthened to ensure
that the acquis is implemented and enforced effectively.
```

English sentence after the POS tagging:

```
administrative+JJ capacity+NN in+IN different+JJ area+NN_NNS need+VB_VBZ to+TO
be+VB strengthen+VB_VBN to+TO ensure+VB that+IN the+DT acquis+NN be+VB_VBZ
implement+VB_VBN and+CC enforce+VB_VBN effectively+RB .+.
```

English sentence after the dependency parsing:

```
1   administrative  administrative  JJ   JJ   _  2   NMOD
2   capacity        capacity        NN   NN   _  6   SBJ
3   in              in              IN   IN   _  2   ADV
4   different       different       JJ   JJ   _  5   NMOD
5   areas           area            NNS  NNS  _  3   PMOD
6   needs           need            VBZ  VBZ  _  0   ROOT
7   to              to              TO   TO   _  8   VMOD
8   be              be              VB   VB   _  6   OBJ
9   strengthened    strengthen      VBN  VBN  _  8   VC
10  to              to              TO   TO   _  11  VMOD
11  ensure          ensure          VB   VB   _  9   OBJ
12  that            that            IN   IN   _  15  VMOD
13  the             the             DT   DT   _  14  NMOD
14  acquis          acquis          NN   NN   _  15  SBJ
15  is              be              VBZ  VBZ  _  11  OBJ
16  implemented     implement       VBN  VBN  _  15  VC
17  and             and             CC   CC   _  16  CC
18  enforced        enforce         VBN  VBN  _  16  COORD
19  effectively     effectively     RB   RB   _  18  ADV
20  .               .               .    .    _  6   P
```

English sentence after transformations:

administrative+JJ capacity+NN different+JJ area+NN_NNS_in+IN need+VB_VBZ strengthen+VB_VBN_to+TO_be+VB ensure+VB_to+TO that+IN acquis+NN_the+DT implement+VB_VBN_be+VB_VBZ and+CC enforce+VB_VBN effectively+RB .+.

Translated Turkish sentence in its morphological form:

Müktesebat+Noun+Gen etki+Noun+With biçim+Noun+Loc uygula+Verb+Pass+Inf2+P3sg+Acc temin+Noun et+Verb+Inf1+üzere+PostpPCNom ,+Punc değişik+Adj alan+Noun+A3pl+Loc+Rel idari+Adj kapasite+Noun+Gen güç+Noun+Acquire+Caus+Pass+Inf2+P3sg gerek+Verb+Prog2+Cop .+Punc

Translated Turkish sentence in its surface form:

Müktesebatın etkili biçimde uygulanmasını temin etmek üzere, değişik alanlardaki idari kapasitenin güçlendirilmesi gerekmektedir.

Bibliography

Doug J. Arnold, Lorna Balkan, Siety Meijer, Robert L. Humphreys, and Louisa Sadler. *Machine Translation: An Introductory Guide*. Blackwells-NCC, London, 1993.

Charles Barber. *The English Language: A Historical Introduction*. Cambridge University Press, 1999.

Alexandra Birch, Miles Osborne, and Philipp Koehn. CCG supertags in factored translation models. In *Proceedings of SMT Workshop at the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311, 1993.

Sabine Buchholz and Erwin Marsi. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL*, pages 149–164, 2006.

Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of EUROSPEECH*, pages 2707–2710, 1997.

Michael Collins, Philipp Koehn, and Ivona Kucerova. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 531–540, 2005.

İlknur Durgar El-Kahlout. *A Prototype English-Turkish Statistical Machine Translation System*. PhD thesis, Sabancı University, 2009.

İlknur Durgar El-Kahlout and Kemal Oflazer. Initial explorations in English to Turkish statistical machine translation. In *Proceedings of SMT Workshop at the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2006.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech 2008*, pages 1618–1621. ISCA, 2008.

Jan Hajič, Jan Hric, and Vladislav Kuboň. Machine translation of very close languages. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 7–12, 2000.

Dilek Zeynep Hakkani-Tür, Gökhan Tur, Kemal Of lazer, Teruko Mitamura, and Eric H. Nyberg. An English-to-Turkish interlingual MT system. In *Proceedings of AMTA'98: Conference of the Association for Machine Translation in the Americas*, pages 83–94, 1998.

Johan Hall, Jens Nilsson, and Joakim Nivre. English Maltparser model, February 2008. URL `http://w3.msi.vxu.se/users/jha/maltparser/mco/english_parser/engmalt.html`.

Hieu Hoang and Philipp Koehn. Improving mid-range re-ordering using templates of factors. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 372–379, March 2009.

Daniel Jurafksy and James H. Martin. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall, 2000.

Celia Kerslake and Aslı Göksel. *Turkish: A Comprehensive Grammar*. Comprehensive Grammars. Routledge (Taylor and Francis), New York, 2005.

Kevin Knight. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25:607–615, 1999.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, 2005.

Philipp Koehn and Hieu Hoang. Factored translation models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2007.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, 2003.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, pages 177–180, 2007.

Alon Lavie. Machine translation overview, August 2008. URL `http://www.cs.cmu.edu/ãlavie/Presentations/MT-Overview-IC-Aug08.ppt`.

Eugene E. Loos, Susan Anderson, Dwight H. Day Jr., Paul C. Jordan, and J. Douglas Wingate. *Glossary of Linguistic Terms*. LinguaLinks Library. SIL International, 2003.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313–330, 1994.

Sonja Niessen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 2000.

Joakim Nivre, Hall Johan, Nilsson Jens, Chanev Atanas, Gülşen Eryiğit, Sandra Kübler, Marinov Stetoslav, and Erwin Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering Journal*, 13(2): 99–135, 2007.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. Improved alignment models for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, 1999.

Kemal Oflazer. Two-level description of Turkish morphology. In *Proceedings of the 6th conference on European chapter of the Association for Computational Linguistics*, pages 472–472, Morristown, NJ, USA, 1993. Association for Computational Linguistics. ISBN 90-5434-014-2.

Kemal Oflazer. Statistical machine translation into a morphologically complex language. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 376–387, 2008.

Kemal Oflazer and İlknur Durgar El-Kahlout. Exploring different representational units in English-to-Turkish statistical machine translation. In *Proceedings of SMT Workshop at the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 25–32, 2007.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, 2001.

Zeki Sagay. A computer translation from English to Turkish. Master's thesis, METU, 1981.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, 1994.

Andreas Stolcke. SRILM - an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, pages 901–904, 2002.

David Talbot and Miles Osborne. Randomised language modelling for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, pages 252–259, 2003.

Ferhan Türe. A hybrid machine translation system from Turkish to English. Master's thesis, Sabancı University, 2008.

Çiğdem Keyder Turhan. An English to Turkish machine translation system using structural mapping. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 320–323, 1997.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. Using a dependency parser to improve SMT for subject-object-verb languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253, June 2009.