INTELLIGENT HYBRID MAN-MACHINE TRANSLATION EVALUATION

A THESIS

Presented to the Graduate School
Faculty of Engineering, Alexandria University
In Partial Fulfillment of
the Requirements for the Degree

Of
Master of Science

In
Computer and Systems Engineering

By
Ibrahim Ahmed Ibrahim Saleh Sabek

June 2014

Supervisors' Committee: Approved

Prof. Dr. Nagwa Moustafa El-Makky …….

Prof. Dr. Soheir Ahmed Fouad Bassiouny ……..

INTELLIGENT HYBRID MAN-MACHINE TRANSLATION EVALUATION

Presented by

Ibrahim Ahmed Ibrahim Saleh Sabek

For the Degree of Masters of Science

In

Computer and Systems Engineering

Examiners' Committee: Approved

Prof. Dr. Amin Ahmed Shoukry …….
Prof. Dr. Nagwa Moustafa El-Makky ……..
Prof. Dr. Soheir Ahmed Fouad Bassiouny ……..
Prof. Dr. Saleh Abd-Elshakour El-Shahaby ……..

Vice Dean for Graduate Studies and Research

Prof. Dr. Heba Wael Leheta .....

ACKNOWLEDGEMENT

First, I thank ALLAH for giving me the ability to complete the work on this thesis.

I thank my advisors Prof. Dr. Nagwa ElMakky, Prof. Dr. Soheir Bassiouny and Dr. Noha Yousri for their advising and help during working on the thesis. This work could not have reached this phase, if it were not for their support, advice and guidance.

I would like to acknowledge my gratitude for Microsoft Advanced Technology Lab in Cairo (ATLC). We have benefited greatly from their support to run our experiments.

Also, I am deeply grateful for my fellow TAs for their company and the experience I shared with them. In particular, I am grateful to: Ahmed Essam, Mahmoud Fouad and Victor Zakhary for their sincere help and support during my experience as a TA.

Last but not the least, I thank my wife, my brother and my sisters for their continuous support and encouragement.

ABSTRACT

Machine Translation (MT) has grasped a lot of attention in translation communities during the recent years and become a crucial part in almost all search engines. However, the widespread of MT technology depends on the trust associated with its outputs. Different approaches have been introduced to address the issues of evaluating translations from one natural language to another. Automatic metrics have been developed to predict the quality of MT outputs. Although these metrics are efficient in terms of speed, the existence of reference translations is assumed. Another research direction, known as Quality Estimation (QE), was proposed to exploit human assessments for evaluation based on machine learning techniques and without reference translations.

Both of automatic metrics and QE approaches have drawbacks. Automatic metrics paid little attention to capture any information at linguistic levels further than lexical. Therefore, these metrics are considered superficial. On the other hand, QE approaches rely only on human assessments which are much more expensive to obtain. Moreover, human assessments can vary for the same translated sentence.

In this thesis, the drawbacks of these two directions are addressed. We extracted a set of linguistic and data-driven features from parallel corpora to evaluate MT outputs. The advantages of these features are twofold. First, they provide a deep linguistic insight which addresses a key issue in automatic metrics. Second, these features are extracted from parallel corpora without the need for expensive human assessments. The experimental evaluation shows that our proposed system outperforms state-of-the-art automatic metrics in terms of accuracy.

Moreover, if human assessments are available, the proposed approach can benefit from them while solving the inconsistency issues of these assessments. A probabilistic inference model was devised to infer the credibility of human assessments. Trusted human assessments can then be used to improve the accuracy of the proposed system.

CONTENTS

1 Introduction **1**

1.1 General 1

1.2 Motivation of The Work 2

1.3 Objectives and Contributions 3

1.4 Thesis Organization 4

2 Literature Survey **5**

2.1 Introduction 5

2.2 Machine Translation Overview 5

2.2.1 Machine Translation Approaches 6

2.2.1.1 Rule-based Approach 6

2.2.1.2 Empirical-based Approach 8

2.2.1.3 Hybrid Approaches 11

2.2.2 Advantages and Disadvantages of Current MT approaches 12

2.3 Machine Translation Evaluation Survey 12

2.3.1 Human Evaluation of Machine Translation 13

2.3.1.1 Fluency and Adequacy 13

2.3.1.2 Reading Time 14

2.3.1.3 Post-editing Time 14

2.3.2 Automatic Evaluation of Machine Translation 14
2.3.2.1 Traditional MT Evaluation Metrics 15
2.3.2.2 Quality Estimation Approaches 18
2.4 The Need to Extend Related Work 21
2.5 Conclusion 21

3 The Proposed Approach **23**
3.1 Introduction 23
3.2 Contribution Bases 23
3.2.1 Alignment Models 24
3.2.1.1 Problem Definition 24
3.2.1.2 Word-based Alignment 27
3.2.1.3 Phrase-based Alignment 29
3.2.2 Bayesian Inference 31
3.2.2.1 Introduction to Graphs 32
3.2.2.2 Introduction to Factors 32
3.2.2.3 Introduction to Factor Graph Representation 33
3.2.2.4 Infer.NET Framework 34
3.3 Proposed System Architecture 36
3.4 Fetching Translation Matches 40
3.5 Alignment Module 40
3.6 Features Scoring Module 41
3.6.1 Outside Source/Target Alignment Features 41
3.6.2 Inside Source/Target Alignment Features 42
3.6.3 Source/Target Coverage Features 43

3.6.4 Source/Target Statistics Features 43
3.7 Scoring Aggregation Module 44
3.8 Inference Module 45
3.8.1 Probabilistic Inference Model 46
3.8.1.1 Scoring Model 46
3.8.1.2 Prior Estimation 47
3.8.1.3 Adaptation to Ordinal Scores 47
3.8.2 Factor Graph Representation 48
3.9 Complexity Analysis 49
3.10 Conclusion 50

4 Experimental Evaluation **51**
4.1 Evaluation Metrics 51
4.1.1 Correlation with Human Judgment 51
4.1.2 Cumulative Distribution Function (CDF) of Correlation 52
4.1.3 Response Time 52
4.2 Tools and Implementation 53
4.3 Datasets and Human Judgments 53
4.3.1 Europarl Dataset 54
4.3.2 WMT Datasets 55
4.4 Performance Evaluation 56
4.4.1 Effect of N-gram Length 57
4.4.2 Effect of Different Combinations of Features 58
4.4.3 Comparison with Automatic Evaluation Metrics 60
4.4.4 Effect of Weighing Features with Human Assessments 61

4.4.5 Comparison with Quality Estimation Approaches 61

4.5 Conclusion 64

5 Conclusions and Future Work **65**

5.1 Conclusions 65

5.2 Future Work 66

LIST OF FIGURES

2.1 Transfer-based Machine Translation. 7

2.2 Interlingua-based Machine Translation. 8

2.3 Architecture of an Empirical-based Machine Translation. 9

2.4 Example on n-gram matches with the reference translation for the BLEU score. 15

3.1 Example on alignment between German and English sentences 24

3.2 Example on word alignment matrix: Words in the English sentence (rows) are aligned to words in the German sentence (columns) as indicated by the filled points in the matrix 26

3.3 Example on merging source-to-target and target-to-source alignments by taking the intersection (black) or union (gray) of the sets of alignment points 27

3.4 Extracting a phrase from a word alignment 29

3.5 Different cases of phrase pairs to show their consistency with a word alignment 31

3.6 Different examples on the factor graph representation 35

3.7 Inference process through Infer.NET 36

3.8 Proposed System Architecture. Main modules are rounded with the solid line, and optional modules are rounded with the doted line. 37

3.9 Sequence diagram for the alignment operations of the proposed system. . . 38

3.10 Sequence diagram for aggregating scores in the proposed system. 39

3.11 Factor graph of the inference model 49

4.1 Effect of different N-gram lengths on the correlation with human judgments using WMT 2007 datasets. 57

4.2 Effect of different N-gram lengths on the correlation with human judgments using Europarl 2005 datasets. 57

4.3 Effect of different N-gram lengths on the response time using WMT 2007 datasets. 59

4.4 Effect of different N-gram lengths on the response time using Europarl 2005 datasets. 59

4.5 CDF of correlation with human judgments for different combinations of features using WMT 2007 datasets. 60

4.6 CDF of correlation with human judgments for different combinations of features using Europarl 2005 datasets. 60

4.7 Comparison with traditional evaluation metrics using WMT 2007 and WMT 2013 datasets. 62

4.8 Effect of weighing features with normal human scores using WMT 2007 and WMT 2013 datasets. 63

4.9 Effect of weighing features with human confidence scores from the proposed inference model using WMT 2007 and WMT 2013 datasets. 63

4.10 Comparison with quality estimation approaches using Spanish-to-English dataset from WMT 2013. 64

LIST OF TABLES

4.1 The number of sentences and words in Eurporl 2005 for different European languages. 54

4.2 Statistics of French-to-English and Spanish-to-English datasets from WMT 2007. 56

4.3 Statistics of French-to-English and Spanish-to-English datasets from WMT 2013. 56

4.4 Correlation with human judgments for WMT 2007 and Europarl 2005 datasets using 5-gram configuration. 58

4.5 Correlation with human judgments for WMT 2007 and Europarl 2005 datasets using 9-gram configuration. 58

4.6 The effect of different combinations of features classes on correlation with human judgments using WMT 2007 datasets. 61

4.7 The effect of different combinations of features classes on correlation with human judgments using Europarl 2005 datasets. 61

LIST OF ALGORITHMS

1 Extracting consistent phrases with word alignment A 30
2 Phrases extraction function $extract(s_{start}, s_{end}, t_{start}, t_{end})$ 30

CHAPTER 1

INTRODUCTION

1.1 General

Machine translation evaluation is an important research field that aims to assess translations that are generated from machine translation systems with respect to many aspects such as running time, complexity and translation quality. In recent years, many applications widely translate the featured content and make it available on the internet. This poses new challenges as well as opportunities for research efforts in machine translation evaluation.

Machine translation is the process of using software to translate text from one natural language to another. During the last decade, the rapid development of the Internet raised the interest in machine translation to overcome the barrier of language. Examples of machine translation-based domains include search engines, social networks, data mining, recommendation systems and so on. Effective evaluation approaches should be concerned with all translation aspects. However, the most important aspect of machine translation is the output quality. This motivated an extensive research effort in the area of evaluating the quality of translation texts.

Machine translation evaluation methods can be divided into two main categories: *Automatic* evaluation metrics and *Quality Estimation* (QE) approaches. Automatic evaluation metrics are techniques that do not include any human interaction (totally unsupervised). Such metrics can be used for applications that require quick, coarse-based translation evaluation. However, this type of metrics cannot accurately mimic all human evaluation aspects and require reference translations. In contrast, QE approaches make use of previous human assessments to predict the quality of unseen translations using machine learning algorithms

without need to reference translations. However, obtaining human assessments is expensive. Moreover, these human assessments can be inconsistent. This thesis addresses the drawbacks of both automatic metrics and quality estimation approaches. The objective is to build a system that draws upon the research in both fields.

1.2 Motivation of The Work

The last decade has witnessed extensive research effort in designing automatic evaluation metrics for machine translation. These metrics compare automatic translations against reference translations produced by humans. Research in these metrics has concentrated only on the lexical level. Although linguistic and data-driven features have widespread enhancements in MT outputs, no satisfactory investigation of these features in the evaluation field has been provided. The success of linguistic features in delivering accurate MT outputs inspired us to explore them in evaluating machine translation outputs.

Recently, post-editing of MT outputs has attracted considerable interest among human translators. In post-editing, the output of MT is improved to ensure that it meets high level of quality. Automatic metrics need reference translations to evaluate the quality of translations to post-edited. However, in practice, reference translations can hardly be available. This problem was addressed using QE approaches where previous human assessments were exploited by machine learning techniques to predict the quality of translations without reference translations. However, human assessments are much more expensive to obtain than a parallel corpus of source and target sentences. In addition, these assessments can vary for the same translation. This motivated us to find a solution that can rely on parallel corpora rather than human assessments. In this thesis, we propose a set of linguistic and data-driven features extracted from parallel corpora to evaluate machine translation. On the other hand, we might benefit greatly from credible human assessments if these assessments are available. This stimulated us to provide a probabilistic inference model to infer the credibility of these assessments. The proposed approach has the advantage of evaluating a whole batch of translations efficiently which is a basic need for large-scale applications.

1.3 Objectives and Contributions

This thesis aims at introducing a hybrid machine translation evaluation approach that addresses the drawbacks of automatic metrics and benefits from human assessments if these assessments are available. The contributions of our work in this thesis can be summarized as follows:

1. Proposing a novel set of linguistic and data-driven features for evaluating translations that make use of alignment-based models and statistics built from offline parallel corpora. These features are extracted on different granularity levels (e.g. phrases or subphrases) with an overall objective of improving the evaluation accuracy.

2. Proposing a novel inference model that infers the credibility of available human assessments. The inference model addresses the high variations among human scores by learning uncertainties in these scores and identifying bad judgments to be discarded or re-examined. Trusted human judgments can then be used to enhance the accuracy of the proposed system.

3. Proposing an aggregation scoring formula to integrate scores from linguistic and data-driven features with confidence scores from trusted human assessments. This formula provides an opportunity to harness extracted features and credible human judgments for a unified final evaluation score.

4. Evaluating the proposed approach using well known datasets and comparing its performance to traditional machine translation evaluation metrics, and state-of-the-art quality estimation approaches.

1.4 Thesis Organization

The rest of the thesis is organized as follows: Chapter 2 gives a background about machine translation and an overview of related work in the area of machine translation evaluation. Chapter 3 presents the details of the proposed approach. Chapter 4 presents the results of the conducted experiments. Chapter 5 provides the conclusions of this work, and discusses possible future directions.

CHAPTER 2

LITERATURE SURVEY

2.1 Introduction

In this chapter, a survey on machine translation approaches and evaluation is given. In section 2.2, an overview of machine translation is presented. A detailed survey of automatic machine translation evaluation metrics and quality estimation approaches is given in section 2.3. The need to extend the related work is shown in section 2.4. Finally, the chapter is concluded in section 2.5.

2.2 Machine Translation Overview

The conversion from a language into another efficiently and quickly has become a common concern for humanity, because language is the most significant means for human to communicate. From another prescriptive, the rapid development of the Internet in the international community requires to overcome the barrier of language [1]. Translation by people can't meet the society needs of large contents, so the use of machine to automate the translation process has become an urgent need.

Machine Translation is a sub-field in the intersection area between computational linguistics and nagtural language processing that investigates how to use the computer in the conversion process from a natural source language into another natural target language. It has been widely applied to numerous application domains such as search engines, social networks, data mining, recommendation systems, etc. The main challenge faced by machine translation is to satisfy certain quality requirements while maintaining the computation com-

plexity of the approach to a minimum.

Text collections are called corpora, and for machine translation we are especially interested in parallel corpora, which are texts, paired with a translation into another language. Texts differ in style and topic, for instance transcripts of parliamentary speeches versus news wire reports. Preparing parallel texts for the purpose of machine translation may require crawling the web, extracting the text from formats such as HTML, as well as document and sentence alignment [2].

2.2.1 Machine Translation Approaches

There has been a significant research effort on machine translation approaches. In this subsection, we review some of the machine translation approaches from literature. The machine translation process can be simplified to three stages: the analysis of source-language text, the transformation from source-language to target-language text, and the target-language generation.

Machine translation approaches can be divided into different categories. First, there are rule-based translation approaches in which translation knowledge base consists of dictionaries and grammar rules is used. Rule-based translation systems can be divided into three catalogs: literal translation method, interlingua-based method and transfer-based method. According to [3], there are also empirical-based methods that aim at using knowledge base as the core. In general, such knowledge base consists of parallel corpuses on different text levels e.g. sentences, phrases, words, etc. Traditionally, empirical-based translation approach can be divided into two different classes: the statistic-based translation and the example-based translation approach. It should be noted that there are approaches that combine more than one category. For example, METIS-II [4] considers both example-based, and rule-based machine translation systems. In the following subsections, an overview of different machine translation approaches from different categories are discussed.

2.2.1.1 Rule-based Approach

Literal translation method: It is a simple form of rule-based machine translation. Literal translation is called direct translation, word-based translation or dictionary-based translation.

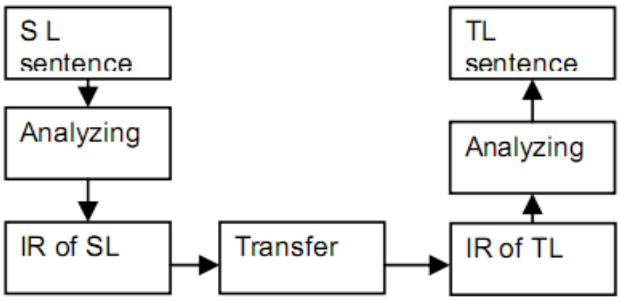


Figure 2.1: Transfer-based Machine Translation.

The basic idea is that the words will be translated word by word, usually without much consideration for context match between them [5]. As an example, it basically works as follows: a word or sentence from the source language is selected, and then looked up in the dictionary for the corresponding word or sentence in the target language. That is why the literal translation is generally designed for a particular language pair and it is not versatile.

IBM701 is the earliest literal machine translation system and it was introduced to the world in 1954. However, Systran [6] is considered the most popular literal translation system till now. At the beginning, it was adopted to translate only from Russian to English. Later it was extended to support different language pairs. Systran [6] has a great impact on the machine translation development. Due to the simplicity of the literal translation techniques, they do not have unsatisfactory results in general, and they can work effectively only for simple translation tasks. Thus, other translation techniques were developed to provide better results.

Transfer-based method: Along with the development of the literal translation method, the transfer-based method was proposed. The transfer-based method performs an analysis of the sentence structure and generates the target-language text based on the different linguistic rules of the different languages. In the transfer-based method, the translation correspondence is word-to-word, and we have there dictionaries: the source dictionary, the source-target bilingual dictionary and the target dictionary.

As shown in figure 2.1, the translation starts with analyzing the source text for syntax, semantics and morphology to create and internal representation. Then, we use this representation as well as both of the grammatical rules and bilingual dictionaries to generate the translation. TAUM [7] and METEO [8] are examples of transfer-based method.

Interlingua-based method: In parallel with the development of literal translation method

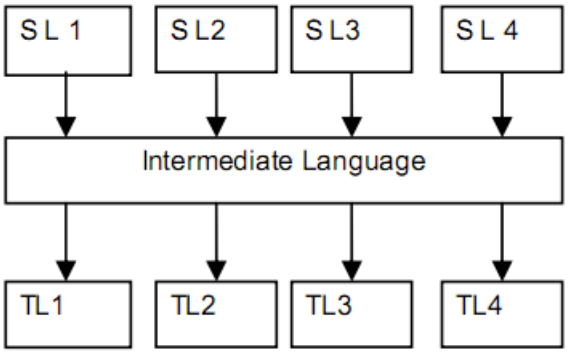


Figure 2.2: Interlingua-based Machine Translation.

and transfer-based method, the Interlingua-based method came into being [9], which can be considered a better alternative, specially when it is compared to both literal and transfer-based methods. As shown in figure 2.2, the process consists of two phases:

- Analyzing and converting the source language into an Interlingua, which is an abstract language-independent representation and can be applied to all languages.
- Converting the Interlinuga into the target language.

There are two advantages of the Interlingua-based machine translation as stated in [10]. First, this approach can localize the development of machine translation system. In other words, to develop a machine translation system for a certain language, you need to collect expert people to analyze and generate the rules for a certain language. Also, native speakers are supposed to develop dictionaries. Using the Interlingua interface, we can completely separate the analysis and generation of rules from the remaining machine translation stuff. Developers of machine translation systems can proceed independently from the human language experts. Developers need only to know the Interlingua and the language being generated. The second advantage is that knowledge described in Interlingua may be used by the analysis systems for each language. This knowledge is essential for high quality machine translation.

2.2.1.2 Empirical-based Approach

The main drawback of the rule-based approach is the high cost as it depends mainly on human experts to specify a set of rules aimed at describing the translation process. Typically,

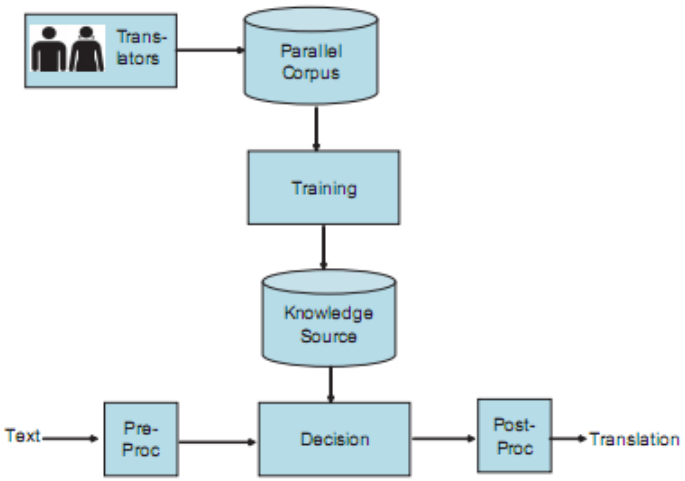


Figure 2.3: Architecture of an Empirical-based Machine Translation.

linguistic experts are more expensive and hard to find.

Empirical-based approaches aim at exploiting example translation to compute automatically the knowledge sources required to develop machine translation systems. The main merit of empirical approaches is that machine translation systems for new language pairs and domains can be developed very quickly, provided that sufficient training data is available. Figure 2.3 shows the architecture of an empirical MT system.

The Empirical-based machine translation can be divided into two main classes, the statistic-based machine translation and the example-based machine translation. In this subsection, a brief summary of the two empirical-based methods is provided.

Statistics-based method: The main idea is that any sentence in a language has a probability to be the translation of a sentence in another language. However, this probability varies and needs to be maximized. The more we find the maximum probability sentence, the more accurate machine translation becomes. So, statistic-based machine translation is concerned with three main problems: modeling problems, training problems and decoding problems. The main purpose of the modeling problem is to define the calculation method of the probability of sentence translation from the source language to the target language. Training problem focuses on how to exploit the corpus to estimate the parameters of this model. Finally, the decoding problem is to find the maximum probability translation for any sentence of source language based on the known models and parameters.

The researchers in IBM proposed the basic mathematical model of statistical machine

translation which considers a linear relationship between words only, and ignores the structure of the sentence [11]. This can lead to low quality translation if the word order of the two languages is totally different. To avoid this limitation, the syntactic structure or semantic structure was considered besides the language and translation models.

Statistic-based machine translation itself has been evolved from the linear model to log-linear model. From another perspective, the main statistical machine translation model has been developed from early word-based model to phrase-based model and syntax-based statistical translation model. Moses [12] is known as the most popular open-source toolkit for statistical machine translation using the log-linear model. The purpose of this toolkit is to allow training translation models for any language pair automatically. Once the trained model is built, an efficient search algorithm can quickly find the highest probability translation among the exponential number of choices.

Google and Bing Translators are the most popular examples of the statistics-based machine translation. The main idea behind them is to prepare a parallel corpus from the bilingual web content. When a source language sentence is submitted, the most common corresponding word is selected based on the computed probability values, and finally gives the translation results. The accuracy of the translation model and estimated model parameters relies directly on the size of the parallel corpus. Large-scale bilingual corpora can refine the translation quality. Also, the probability model coverage of the parallel corpus is an important factor that can increase the overall translation accuracy because it selects the translation text based on the computer statistics from the parallel corpus.

Statistics-based machine translation still suffers from some limitations. First of all, it doesn't take into account the semantic information of the translations. This conclusion was confirmed by experimental results shown in [13]. Another problem was proposed by Chomsky in [13] about its inefficiency in dealing with the problem of long-distance between subject and verb, a famous problem in natural language processing.

Example-Based method: Example-based machine translation simulates the human translation process. It consists of three stages. First, and similar to human minds, it breaks the source language text into sentences, then decomposes these sentences into smaller phrases. The purpose of translating these different parts is to compose these fragments into one long sentence again. Second, it translates the source phrases into target phrases by analogy. Third, it combines the resulting phrases into sentences.

The translation process depends on a bilingual alignment between phrases from source and target languages. Given a source sentence as an input, the system searches for the most similar sentence from the source language base, then according to the translation of target language, some words or phrases in target language will be changed and the translation of target language will be done. The principle of this kind of translation system is proposed in the first example-based machine translation system in [14], which uses large number of bilingual examples as the backbone of the system.

To have a successful Example-based machine translation, it is required to build a large bilingual example corpus. For the same input sentence, using large-scale corpus increases the translation results because it achieves higher matching rate. In addition, there are some problems that should be solved such as how to define the measurement of similarity between the input and examples in the corpus. This problem should be handled carefully because selecting irrelevant examples to use will low the overall translation score and hence leads to low translation quality. Besides, the problem of alignment of bilingual text should be considered. Sentence alignment is not the only requirement, most of time the phrase alignment or even lexical alignment is also needed. Due to these limitations, few translation systems rely only on Example-based approach. However, it is used as one of the multiple translation engines, to improve the overall translation quality.

2.2.1.3 Hybrid Approaches

The machine translation systems discussed above either used traditional empirical-based or rule-based approaches. A traditional rule-based machine translation system, as in [15] and [16], can easily provide inconsistencies, and it is too rigid to be robust. In contrast, the empirical-based approach [17] is robust in handling ill formed sentences. However, the running time for processing long sentences is affected by the number of words in a sentence. Increasing the number of words will increase the running time significantly. In addition, the overall translation quality depends on the quality of collected examples in the parallel corpus. The translation accuracy increases when the matched units are phrases and not on the word-level.

To avoid limitations of both methods, hybrid approaches were proposed to combine the best features of both methods. A lot of current research in machine translation is neither

based purely on linguistic knowledge nor on statistics, but includes some degree of hybridization. Currently the research in this field is directed at the development of hybrid MT systems which integrate more than one approach to MT, the idea being that integration will help achieve properties that combine the advantages of the approaches involved. Lingstat [18] is a hybrid MT system, combining statistical and linguistic techniques while METIS-II [4] is a hybrid machine translation system, in which insights from statistical, example-based, and rule-based machine translation are used. Cunei [19] is another well-known example on the hybrid approach where a joint model of statistics-based and example-based MT was proposed. This system was developed in Carnegie Mellon University based on linguistic and data-driven features extracted from parallel corpora.

2.2.2 Advantages and Disadvantages of Current MT approaches

All the current machine translation approaches have advantages and disadvantages. As in the scope of rationalism, the literal translation method, the transfer-based method, and the interlingua-based method are rule-based approaches. The typical disadvantage is that the grain size is too large, that is, the computer language can not fully describe the actual infinite rules.

Statistic-based methods and example-based methods are empirical-based approaches with a typical disadvantage of data sparseness. In other words, because of infinite languages, any high-performance computer can not count all usages of the phrase. With the disadvantages of these approaches, more and more approaches appear to integrate different machine translation approaches together.

2.3 Machine Translation Evaluation Survey

Machine Translation evaluation has been a very attractive field for research during the last decade. A lot of metrics and approaches have been proposed to deal with the translation quality and running time complexity. These approaches have tried to consider all challenges of machine translation using different techniques. In this section, we present in brief the related work in machine translation evaluation.

2.3.1 Human Evaluation of Machine Translation

We summarize the most important aspects considered by humans to manually evaluate the output of machine translation systems. These aspects are widely used as challenges for evaluating the performance of different MT systems.

2.3.1.1 Fluency and Adequacy

Fluency is defined as the degree to which the translation is well-formed according to the grammar of the target language. [20] proposed a set of methods to measure fluency by focusing on specific syntactic constructions such as relative clauses, aligned sentences, etc. From another perspective, others simply ask judges to provide rating for the whole sentence on a n-point scale. Commonly, they used the following five point scale: *a*) 5 points for flawless level; *b*) 4 points for good level; *c*) 3 points for non-native level; *d*) 2 points for disfluent level and *e*) 1 point for incomprehensible level. Other work proposed to automatically measure the complexity of the generated target language text against a language model derived from ideal translations.

Adequacy is defined as the quantity of the information existent in the original text that a translation contains. Adequacy also has a similar scale to fluency as follows: *a*) 5 points for All; *b*) 4 points for Most; *c*) 3 points for Much; *d*) 2 points for Little and *e*) 1 point for None.

It is obvious that fluency and adequacy are different aspects of evaluation. For example, a translation might be disfluent but contain all the information from the source. Thus, separate scales are needed to measure these different flavors. However, in practice, it seems that people mix these two aspects of translation. Also, there are no specific guidelines to rate translations such as how many grammatical errors separate the different levels of fluency or how to quantify the amount of information to distinguish between different adequacy levels. This leads to a subjective assessment for each individual which might be inaccurate in most cases.

2.3.1.2 Reading Time

Reading time evaluates the closeness between the Words Per Minute (WPM) rate of the generated text and the WPM rate of natural language. The higher WPM rate is, the higher the quality of translation becomes. It targets large-scale machine translations. There are two types of reading time:

- Oral reading time: for each document, the evaluators should read out loud the first paragraph and count the time it takes. The number of words is then used to calculate the WPM rate.
- Closed reading time: as for oral reading time, the WPM needs to be calculated. This is done in the same way. The level of understanding of the readers also needs to be checked to see if it is sufficient. For this check, the reader was requested to answer some basic questions about the text.

2.3.1.3 Post-editing Time

Post-editing time measures how long it is required to transfer generated translations into an acceptable text. Higher values for this measure mean inaccurate translations. Usually, the measured time is normalized by the number of words in the text and multiplied by a fixed scale to avoid too small scores. However, there are two main drawbacks to use this measure. First, it is difficult to specify the nature of all errors and needed time to correct them. Second, it depends on the skill of judges, i.e. some correctors work faster than others.

2.3.2 Automatic Evaluation of Machine Translation

As shown in the previous section, manual evaluation has a lot of advantages and can evaluate many aspects of successful translations. However, manual evaluation is discouraged because human resources are more expensive specially language experts. Here comes the need for automatic evaluation where translations are compared with reference sentences produced by human. The main advantage of automatic evaluation is the re-usability of the algorithm used for every source text, while on the contrary manual evaluation techniques require consider-

Figure 2.4: Example on n-gram matches with the reference translation for the BLEU score.

able time and people. However, automatic evaluation metrics are superficial and can't cover all human evaluation aspects [21].

Effective automatic evaluation metric has to satisfy some requirements as stated in [22]. First of all, a good metric should be as sensitive as possible to differences in MT quality between different systems, and between different versions of the same system. Furthermore, the metric should be consistent (same MT system on similar texts should produce similar scores), reliable (MT systems that score similarly can be trusted to perform similarly) and general (applicable to different MT tasks in a wide range of domains and scenarios).

2.3.2.1 Traditional MT Evaluation Metrics

In this section, traditional automatic metrics are presented in details. These metrics are considered benchmarks for the assessment of any new evaluation metric.

Bilingual Evaluation Understudy (BLEU): is the most popular metric for automatic machine translation evaluation. It was proposed by Kishore Papineni and others in 2001 [23]. The main idea of BLEU is to consider matches of larger n-grams between the input and reference translations. It also handles the role of word order.

Figure 2.4 provides an example on n-gram matches for two systems with a reference translation. System A matches are a 2-gram match for *Arabic officials* and a 1-gram match for *airport*. For system B, *airport security* is a 2-gram match and *Arabic officials are responsible* is a 4-gram match. Given the n-gram matches, we compute n-gram precision, i.e., the ratio of correct n-grams of a certain order n in relation to the total number of generated n-grams of that order:

- System A: 1-gram precision 3/6, 2-gram precision 1/5, 3-gram precision 0/4, 4-gram precision 0/3.

- System B: 1-gram precision 6/6, 2-gram precision 4/5, 3-gram precision 2/4, 4-gram precision 1/3.

To compute the BLUE score, we average the precision of these matches between input and reference translations. Therefore, in our example, system B has higher BLUE score than system A. Usually, multiple reference translations are used while calculating the BLEU score. Given the variability in translation, it is harsh to require matches of the system output against a single human reference translation. If multiple human reference translations are used, it is more likely that all acceptable translations of ambiguous parts of the sentences show up. The use of multiple reference translations works as follows. If an n-gram in the output has a match in any of the reference translations, it is counted as correct. If an n-gram occurs multiple times in the output (for instance the English world the often shows up repeatedly), it has to occur in a single reference translation the same number of times for all occurrences to be marked as correct. If reference translations have fewer occurrences of the n-gram, it is marked as correct only that many times.

BLEU was presented as an alternative for human evaluation that can be used when quick and frequent evaluations are required. However, BLEU doesn't count near matches. In our example, although the *responsibility of* n-gram is not a wrong translation, BLEU doesn't count this n-gram for system A. Moreover, BLEU is known to perform poorly (i.e. not agree with human judgments of translation quality) when evaluating the output of commercial systems like Systran [6], or even when evaluating human-aided translation against machine translation [24]. It has been shown in [25] that BLEU systematically underestimates the quality of rule-based MT systems.

Metric for Evaluation of Translation with Explicit ORdering (METEOR): is a machine translation evaluation metric developed at Carnegie Mellon University [22]. METEOR was designed to explicitly address the weaknesses in BLEU [23]. The main disadvantage of BLEU is that it gives no credit to near matches. One possible solution is to reduce words to their stems before applying metrics. Another way to detect near matches is using synonyms, or semantically closely related words. METEOR proposed the use of stemming and synonyms, along with the standard exact word matching. First, an alignment is performed between unigrams of the input and reference translations. The unigrams in input and reference translations are stemmed to their roots and then backed off to semantic classes. METEOR assigns a score equal to the harmonic mean of unigram precision (that is, the proportion of

matched unigrams out of the total number of unigrams in the evaluated translation) and unigram recall (that is, the proportion of matched unigrams out of the total number of unigrams in the reference translation).

Given a pair of translations to be compared (a candidate translation string and a reference translation string), METEOR creates an alignment between the two strings as the first step. An alignment is a mapping between unigrams, such that every unigram in the candidate translation must map to zero or one unigram in the reference translation, and to no unigrams in the same string. Thus in a given alignment, a single unigram in one string cannot map to more than one unigram in the other string.

This alignment is incrementally produced through a series of stages, each stage consisting of two distinct phases. In the first phase an external module lists all the possible unigram mappings between the two strings. In the second phase of each stage, the largest subset of these unigram mappings is selected such that the resulting set constitutes an alignment as defined above. If more than one subset constitutes an alignment, and also has the same cardinality as the largest set, METEOR selects that set that has the least number of unigram mapping crosses.

The main drawback of METEOR is that its method and formula for computing a score is much more complicated than BLEU's. The matching process involves computationally expensive word alignment. There are many more parameters such as the relative weight of recall to precision, the weight for stemming or synonym matches that have to be tuned.

General Text Matcher (GTM): GTM [26] is based on accuracy measures as precision, recall and F-Measure. GTM mainly measures the overlap between strings, rather than overlap between bags of items. An exponent parameter is used to weight the size of matching between candidate and reference translations. This parameter controls the relative importance of word order. A value of 1.0 for this exponent parameter reduces GTM to ordinary unigram overlap, with higher values emphasizing order.

In parallel, the authors showed that the correlation between human judgments of MT quality was surprisingly low because of inconsistency issues. Also, the correlation between human judges and all automatic measures of MT quality was quite low.

Translate Error Rate (TER): represents the number of edits needed to change a hypothesis in one of the references, normalized on the length of the references. It was proposed by

Snover and Dorr in [27]. Possible edits include the insertion, deletion, substitution of single words and shifts of word sequence. A similar measure is Word Error Rate (WER) which is expressed as the minimum edit distance between hypothesis and reference at word level. TER is different from WER because it treats shifts of contiguous multi-word sequences as a single operation.

Other Automatic Evaluation Metrics: Many other automatic evaluation metrics are based on comparing automatic translations against human references. Examples of these metrics are NIST [28], ROUGE [29] and Orange [30]. Such comparisons consider lexical information and do not capture all linguistic knowledge incorporated in MT systems. [31] proposed abstract linguistic features to evaluate MT output as a classification problem. [32, 33, 34] showed that metrics incorporating deep linguistic information are robust compared with lexical-based metrics. [35] defined feature functions in a practical way to capture linguistic and contextual information in translations. [36] provided an engineering solution for selecting the best set of scoring features. However, modeling joint dependencies between features is problematic. For example, adding a binary feature will double the size of the feature space. In addition, parameter tuning fails when we extract more than a few dozen of features as stated in [37]. The main observation in these approaches is the greedy nature of integrating all available features which results in low accuracy if there are too many cross-dependent features.

2.3.2.2 Quality Estimation Approaches

The traditional MT evaluation metrics require reference translations in order to measure a score reflecting some aspects of translation quality. Reference translations are usually offered by human efforts. However, in practice, there is usually no golden reference for the translated documents, especially on the internet works [38]. This raises a challenge of evaluating the quality of automatically translated documents or sentences without using reference translations.

Quality Estimation (QE) has recently grasped the attention of professional readers and translators as the main users of MT systems, because it evaluates the quality of unseen translations using machine learning techniques that exploit human assessments obtained for similar previously stored translations instead of using reference translations. However, the main

drawback of QE approaches is that these human assessments are expensive and hard to obtain. In this section, we highlight the main QE approaches proposed till now.

TrustRank: is a ranking algorithm that was proposed in [39] which takes advantage of a supervised machine learning approach (e.g. regression) to decide the suitability of a translation to be published as is or not. Automatic labels are generated using BLEU scores instead of manual annotation for every document in the training set. Using TrustRank, the user is enabled to set a quality threshold and control over the quality of the translations.

Post-editing Objective Annotations: While there were many trials for exploiting human assessment to obtain better accuracy, they come at a prohibitively high cost, mostly in the form of extensive sentences annotation and labeling for different sentence parts. [40] needs labeled dependency structure of the sentence to score it. However, [41] proposed an effective approach to filter out sentences that need high effort for post-editing. The authors use three different types of annotation (post-editing time, post-editing distance and post-editing effort scores) to replace human annotations. Their experiments showed that these annotations can reliably estimate translation quality and post-editing effort for newly coming translations.

Translation Recommendation for Post-editing: a framework was proposed in [42] to select the best translation to post-edit among options from multiple MT and/or translation memory systems. The authors adapted an SVM binary classifier as the framework core, and exploited automatic MT evaluation metrics to approximate human judgments in their experiments.

Predicting Machine Translation Adequacy: an approach was proposed in [43] to inform readers of the target language about the adequacy of translations. This approach was based on human assessments for adequacy and a number of translation quality indicators to contrast the source and translation texts. Experiments with Arabic-English translations showed that the proposed prediction model can yield more reliable adequacy estimators for new translations.

The FBK-UEDIN Quality Estimation System: a system was proposed to explore a set of features extracted from MT engine resources [44] including n-best candidate translation lists. In addition, automatic MT evaluation metrics were used as features. It was designed to predict the required time and effort to perform sentence-level post-editing.

To overcome the problem of needed reference translations for automatic metrics, three

similar MT systems are built and used to provide pseudo-references from which automatic MT evaluation metrics could be computed and used as features. This system was the best winner in the WMT 2013 [45] shared task for quality estimation.

The CNGL Quality Estimation System: a language-independent framework was proposed to predict the quality of sentence-level machine translation [46]. The authors introduced referential translation machines (RTM) for quality estimation of translation outputs. These machines select common training data relevant and close to both the training set and the test set where the selected relevant set of instances are called the interpretants. These interpretants are used to extract features used for measuring the closeness of a given test sentence to the training data and the difficulty of translating this sentence.

RTMs remove the need to access any MT system specific information or prior knowledge of the training data or models used when generating the translations. This system was able to achieve the second best performance according to the official results of the shared task for quality estimation in WMT 2013 [45].

The CMU Quality Estimation System: CMU proposed a quality estimation system trained on features extracted from language models, length statistics of source/target sentences and n-best lists of translation candidates [47].

The authors discussed that the way sentences are translated from one language to another might differ depending on how complex the information is which in turn might be related to the sentence length for both source and target languages. As a simple way of capturing this phenomenon, the parallel training corpus was divided into three classes (short, medium, long) by the length of the source and target sentences. As features for these classes, a binary function was used to indicate the membership of source/target sentences to each class. The prediction models were trained using different classifiers in the Weka toolkit [48]: linear regression, M5P trees, multi layer perceptron and SVM regression.

This system showed competitive results and achieved the third place in the WMT 2013 shared task of quality estimation [49].

Other Work Related to Quality Estimation: there were other proposals that use machine translation techniques and do not need reference translations. So, we survey them under the quality estimation approaches umbrella. For example, machine learning techniques are used in [50] and [51] to filter out translations that need high effort for post-editing on

the sentence level. This work was found useful for small-scale post-editing applications. However, commercial applications require large-scale transactions of post-editing operations. Another research direction [52] focused on the correlation between automated and human assessments to predict the quality of machine translation systems.

A considerable amount of work has been done to provide MT evaluation without reference translations, based on regression [53] and classification [54]. Also, the work in [55] exploits human assessments to evaluate the performance of a set of MT systems.

2.4 The Need to Extend Related Work

Section 2.3 showed that machine translation evaluation techniques can be divided into two main categories: automatic evaluation metrics and quality estimation approaches. However, existing techniques in these two categories suffer from many shortcomings which can be summarized as follows:

- Automatic evaluation metrics are an essential tool for system development of machine translation systems, but not fully suited to computing scores that allow us to rank systems of different types against each other. Developing evaluation metrics for this purpose is still an open challenge to the research community.
- Quality estimation is generally addressed as a supervised machine learning task that depends heavily on human assessments of previous translations. Most of the research work lies on deciding which aspects of human assessment are more valuable for quality estimation and designing feature extractors for them. While simple scoring features can be easily extracted based on human assessments, these assessments are hard to find. Moreover, human assessments are usually inconsistent. In contrast, our proposed approach can benefit greatly from linguistic and data-driven features that can be extracted from parallel corpora which are available and easy to maintain.

2.5 Conclusion

In this chapter, machine translation methods are presented in brief. Then, current machine translation evaluation approaches are surveyed. Finally, the need to extend the current ap-

proaches is discussed. In the next chapter, the proposed approach for hybrid man-machine evaluation of machine translation will be presented.

CHAPTER 3

THE PROPOSED APPROACH

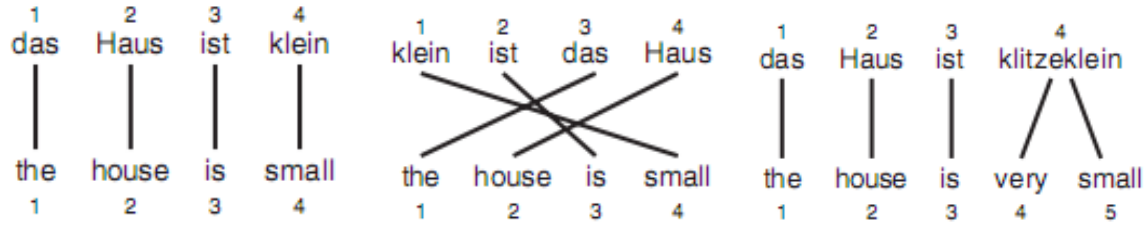
3.1 Introduction

In this chapter, the proposed approach is presented in details. The objective of this work is to provide a highly accurate machine translation evaluation approach that addresses the drawbacks of automatic metrics and benefits from human assessments if these assessments are available.

First, details of alignment models and Bayesian inference as foundations of the proposed approach are given in section 3.2. A brief overview of the proposed system is presented in section 3.3. The process of fetching translation matches is described in section 3.4. Then, the details of proposed system modules are provided. The alignment module is presented in section 3.5, the features scoring module is given in section 3.6, the scoring aggregation module is described in section 3.7, and finally the inference module is given in section 3.8. The complexity analysis of the proposed approach is provided in section 3.9. Finally, the chapter is concluded in section 3.10.

3.2 Contribution Bases

The theoretical foundations upon which the proposed approach is built are explored in this section. These foundations are divided into two parts. First, the theoretical bases of alignment models are presented in section 3.2.1. Second, the fundamentals of Bayesian inference are illustrated in section 3.2.2.



(a) Simple alignment (b) Complex alignment (c) One-to-many alignment

Figure 3.1: Example on alignment between German and English sentences

3.2.1 Alignment Models

An alignment model is defined to determine correspondences between the words/phrases in a sentence in one language with the words/phrases in a sentence with the same meaning in a different language. This model forms an important part of the translation process, as it is used to produce word-aligned parallel text which is used to initialize machine translation systems. Improving the quality of alignment leads to systems which model translation more accurately and an improved quality of output. The success of alignment models in delivering accurate machine translation outputs was an inspiration to use them in evaluating translations.

In this section, we show the problem definition of the alignment process followed by details of word-based and phrase-based alignment models.

3.2.1.1 Problem Definition

Assume we want to translate the German sentence: *das Haus ist klein*. If it is translated word by word, one possible translation can be: *the house is small*. Figure 3.1(a) shows a simple possible alignment between German words and English words of these sentences.

Formally, the alignment should be defined with an *alignment function a*. This function maps, in our example, each English output word at position j to a German input word at position i where German is the source language and English is the target language:

$$a : j \rightarrow i \tag{3.1}$$

In our example, the alignment function would provide the mappings:

$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\} \tag{3.2}$$

This is a very simple alignment, since the German words and their English counterparts are in exactly the same order. While many languages do indeed have similar word order, the target language may have sentences in a different word order than is possible in the source language as shown in figure 3.1(b). Also, languages may also differ in how many words are necessary to capture the same meaning. Figure 3.1(c) shows an example of one German word that requires two English words for alignment.

Formally, let us define the source sentence S as follows:

$$S = \{s_i : 1 \leq i \leq I\} \tag{3.3}$$

where I is the number of words in S. Similarly, we can define the target sentence T as

$$T = \{t_j : 1 \leq j \leq J\} \tag{3.4}$$

where J is the number of words in T. The probability of aligning a source word s_i into a target word t_j can be modeled with the conditional probability function $p(t_j|s_i)$.

Given an alignment function $a : j \rightarrow i$, we can model the probability of aligning sentence S to sentence T as in [2]:

$$P(T, a|S) = \frac{\varepsilon}{(I+1)^J} \prod_{j=1}^{J} p(t_j|s_{a(j)}) \tag{3.5}$$

The fraction before the product is necessary for normalization. To include the special NULL token, there are actually $I+1$ input words. Hence, there are $(I+1)^J$ different alignments that map $I+1$ input words into J output words. The parameter ε is a normalization constant, so that $P(T, a|S)$ is a proper probability distribution, meaning that the probabilities of all possible target sentences and alignments sum up to one:

$$\sum_{T,a} P(T, a|S) = 1 \tag{3.6}$$

Equation 3.5 presents the basic model to align two sentences together based on their

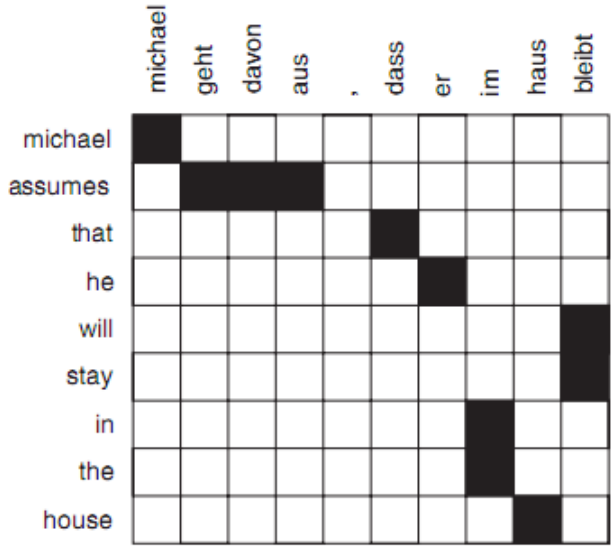


Figure 3.2: Example on word alignment matrix: Words in the English sentence (rows) are aligned to words in the German sentence (columns) as indicated by the filled points in the matrix

lexical words. Having a closer look on this model, many flaws can be observed. The model is very weak in terms of reordering. More often than not, words that follow each other in one language have translations that follow each other in the output language. However, this model treats all possible reorderings as equally likely.

According to [2], IBM proposed more complex models to address the issues of this basic model. The first model added an explicit alignment based on the positions of the input and output words. The second model addressed the fertility issue between different languages. Fertility is the notion that input words produce a specific number of output words in the output language. For example, in most cases, a German word translates to one single English word. However, some German words like *zum* typically translate to two English words, i.e., *to the*. The third model introduced the relative alignment concept where the placement of the translation of an input word is typically based on the placement of the translation of the proceeding input word. However, there is a problem in the third model. In this model, it is possible that multiple output words may be placed in the same position. Of course, this is not possible in practice. This problem was fixed in the fourth model [17].

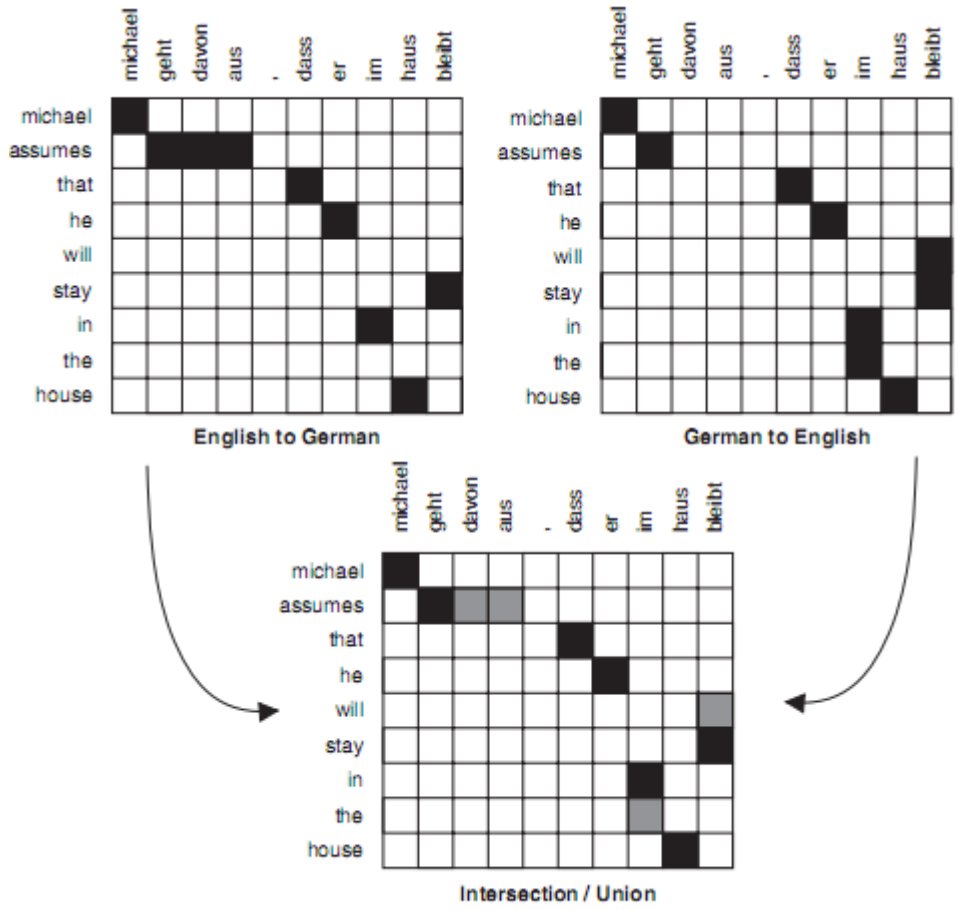


Figure 3.3: Example on merging source-to-target and target-to-source alignments by taking the intersection (black) or union (gray) of the sets of alignment points

3.2.1.2 Word-based Alignment

A word alignment between a sentence and its translation was introduced heavily in the earlier IBM models. One way to visualize the task of word alignment is by a matrix as shown in figure 3.2. Here, alignments between words (for instance between the German *haus* and the English *house* words) are represented by points in the alignment matrix.

Word alignments do not have to be one-to-one. Words may have multiple or no alignment points. As in figure 3.2, the English word *assumes* is aligned to the three German words *geht davon aus*. The German comma is not aligned to any English word. However, it is not always easy to establish what the correct word alignment should be. The main problem is that some function words have no clear equivalent in the other language. This poses a lot of challenges for research efforts in word-based alignment models.

Usually, one word in the target sentence can be ended up with an alignment to multiple words in the source sentence. However, by applying word alignment model as it is, each target word t_j can be exactly aligned to (at most) one source word $s_{a(j)}$ with the probability $p(t_j|s_{a(j)})$. To allow multiple words alignment, we can apply the alignment model in both directions: source-to-target alignment and target-to-source alignment. The two resulting word alignments can then be merged by, for instance, taking the intersection or the union of alignment points of each alignment. This process is called *symmetrization* of word alignments. [2] provided an example on symmetrization using IBM model alignments as shown in figure 3.3. Since these models are not capable of aligning multiple input words to an output word, both a German-to-English and an English-to-German alignment will be faulty. However, these alignments can be merged by taking the intersection (black) or union (gray) of the sets of alignment points.

In this example, the union of the alignments (taking all alignment points that occur in either of the two directional alignments) matches the desired outcome. In practice, this is less often the case, since when dealing with real data, faulty alignment points are established and the union will contain all faulty alignment points. Generally, the intersection will contain reliably good alignment points, but not all of them. The union will contain most of the desired alignment points including additional faulty points. So rather than taking the union or the intersection, the space between these two extremes may be explored. We may want to take all the alignment points in the intersection (which are reliable), and add some of the points from the union (which are the most reliable candidates for additional points). A heuristic has been proposed to exploit the observation that good alignment points neighbor other alignment points. Starting with the alignment points in the intersection, neighboring candidate alignment points from the union are progressively added.

GIZA++: is an implementation of the IBM alignment models [17], and it is commonly used nowadays for word alignment. The code was developed at Johns Hopkins University summer workshop [56] and later refined in [57]. With the help of Expectation-Maximization (EM) [58] algorithm, final word alignment results can be obtained after training the parallel corpus several iterations from two directions (source-to-target direction and vice-versa). The two directions are trained fully independently from each other. Symmetrization is then performed as a post-processing step.

The EM algorithm is employed for the estimation of the parameters of the alignment

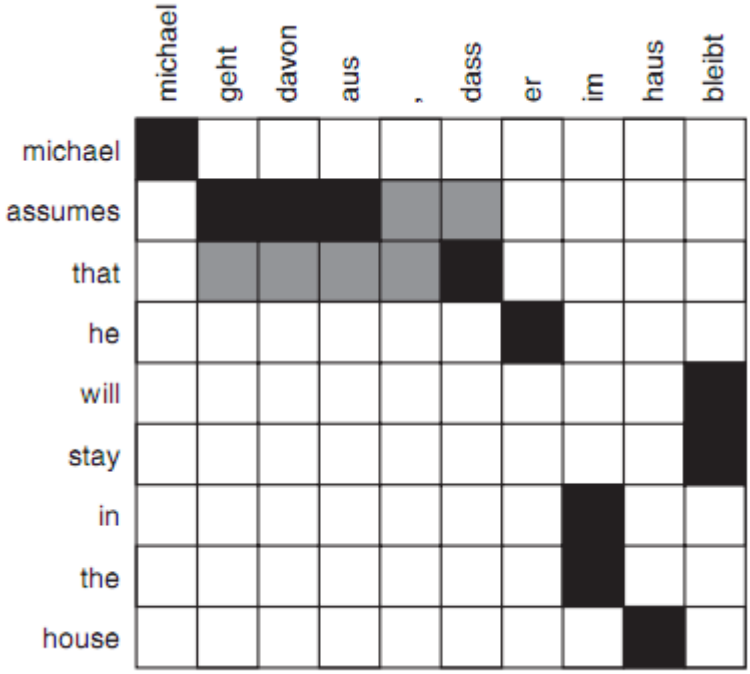


Figure 3.4: Extracting a phrase from a word alignment

models. During the EM algorithm, two steps are applied in each iteration. In the first step, the E-step, the previously computed model or a model with initial values is applied to the data. The expected counts for specific parameters are collected using the probabilities of this model. In the second step, the M-step, these expected counts are taken as fact and used to estimate the probabilities of the next model. A correct implementation of the E-step requires to sum over all possible alignments for one sentence pair.

3.2.1.3 Phrase-based Alignment

The previous section introduced word-based alignment models for machine translation. But words may not be the best candidates for the smallest units for translation. Sometimes one word in a source language translates into two target words, or vice versa. Word-based models often break down in these cases. To acquire phrase-based alignment from a parallel corpus, two steps should be done. First, we create a word-based alignment between each sentence pair of the parallel corpus, and then extract phrase pairs that are consistent with this word alignment.

Consider the word alignment example in figure 3.4, which is the same example used in

Algorithm 1 Extracting consistent phrases with word alignment A

Require: Word alignment A for sentence pair (s,t)
Ensure: Set of phrase pairs BP

```
for t_start = 1 to length(t) do
  for t_end = t_start to length(t) do
    (s_start, s_end) = (length(s), 0)
    for all (s,t) ∈ A do
      if t_start ≤ e ≤ t_end then
        s_start = min(s, s_start)
        s_end = max(s, s_end)
      end if
    end for
    add extract(s_start, s_end, t_start, t_end) to set BP
  end for
end for
```

Algorithm 2 Phrases extraction function $extract(s_{start}, s_{end}, t_{start}, t_{end})$

Require: (s_{start}, s_{end}) in source sentence s, (t_{start}, t_{end}) in target sentence t
Ensure: Set of phrase pairs E

```
valid ← true
if s_end != 0 then
  for all (s,t) ∈ A do
    if t < t_start or t > t_end then
      valid ← false
    end if
  end for
  if valid == true then
    E = {}
    s_s = s_start
    repeat
      s_e = s_end
      repeat
        add phrase pair (s_s..s_e, t_start..t_end) to set E
        s_e++
      until s_e aligned
      s_e--
    until s_s aligned
  end if
end if
return E
```

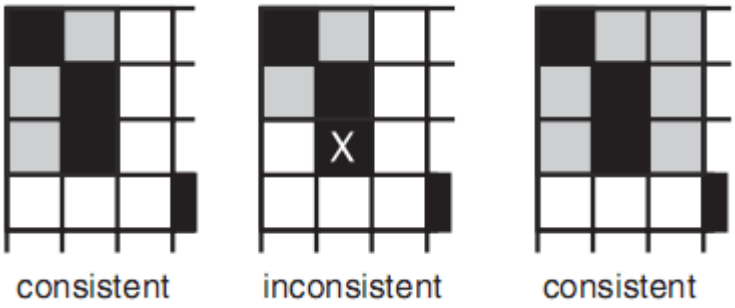


Figure 3.5: Different cases of phrase pairs to show their consistency with a word alignment

the previous section. Given this word alignment we would like to extract phrase pairs that are consistent with it, for example the English phrase *assumes that* and the German phrase *geht davon aus, dass* are aligned, because their words are aligned to each other. Useful phrases for alignment may be shorter or longer than this example. Shorter phrases occur more frequently, so they will more often be applicable to previously unseen sentences. Longer phrases capture more local context and help us to translate larger chunks of text at one time, maybe even occasionally an entire sentence. Hence, when extracting phrase pairs, we want to collect both short and long phrases, since all of them are useful.

Let us assume the following notations: the word alignment matrix is A, the source phrase is $\hat{s} = \{s_i : 1 \leq i \leq n\}$ where n is the number of words in $\hat{s}$, and the target phrase is $\hat{t} = \{t_j : 1 \leq j \leq m\}$ where m is the number of words in $\hat{t}$. We call a phrase pair $(\hat{s}, \hat{t})$ consistent with a word alignment A, if all words in $\hat{s}$ that have alignment points in A have these alignments with words in $\hat{t}$ and vice-versa.

Figure 3.5 shows different cases of phrase pairs to be examined for consistency with a word alignment. The first case is consistent because all words are aligned to each other. The consistency is violated in the second case because one alignment point in the second column is outside the phrase pair. The third case is also consistent in spite of having an unaligned word on the right. The details of phrase extraction are described in algorithms 1 and 2.

3.2.2 Bayesian Inference

Bayesian inference is a method of inference in which Bayes' rule is used to update the probability estimate for a hypothesis as additional evidence is acquired. During the last decade,

expectation propagation [59] and message-passing [60] were proposed as efficient techniques based on *factor graphs* to provide approximate Bayesian inference. A *factor graph* is a particular type of graphical model that enables efficient computation of marginal distributions. Factor graphs have many important success stories in signal processing, cooperative localization, and networking fields.

In this section, we provide some basic definitions of graphs and factors. Then, we give a brief overview of the factor graph representation, followed by an introduction for tools that are used in the proposed work for applying Bayesian inference.

3.2.2.1 Introduction to Graphs

In this section, we will provide some basic definitions about graphs that will be useful in the rest of this work.

Definition 1. ***Graph****: A graph consists of a pair of sets $G(V,E)$ such that $E \subseteq V \times V$, i.e. the elements in E are two-element subsets of V; V represents the set of vertices of the graph, while E denotes the set of edges.*

Definition 2. ***Adjacency****: Given an edge $e \in E$, there are two vertices in V, namely v_1, v_2, such that $e = (v_1, v_2)$. We say that v_1 and v_2 are adjacent and the set of vertices adjacent to vertex v is denoted $N_G(v)$. Given a vertex $v \in V$, there are two edges e_1, e_2 such that $e_1 = (v, v_1)$ and $e_2 = (v, v_2)$ for some $v_1, v_2 \in V$. We say that e_1 and e_2 are adjacent. The set of edges adjacent to edge e is denoted $N_G(e)$.*

Definition 3. ***Bipartite graph****: A bipartite graph is a graph whose vertices can be divided into two disjoint sets U and V such that every edge connects a vertex in U to a vertex in V; that is, U and V are independent sets. Equivalently, a bipartite graph is a graph that does not contain any odd-length cycles.*

3.2.2.2 Introduction to Factors

Through this thesis we deal with functions of several variables. Let $X_1, X_2, ..., X_n$ be a set of variables, in which for each i, X_i takes values in some finite domain D_i. Let $f(X_1, X_2, ..., X_n)$ be a real valued function of these variables, i.e. a function with domain D, where D is as

follows:

$$D = D_1 \times D_2 \times ... \times D_n \tag{3.7}$$

and range the set of real numbers R. The domain D of f is called *configuration space* for the given set of variables $\{X_1, X_2, ..., X_n\}$, and each element of D is a particular configuration of the variables, i.e. an assignment of a value for each input of f. Knowing that the set of real numbers is closed over summation, we will associate *n marginal* functions associated with function $f(X_1, X_2, ..., X_n)$, denoted as $g_{X_i}(x_i)$ for every i. For each $x_i \in D_i$, the value $g_{X_i}(x_i)$ is obtained by summing the value of $f(X_1, X_2, ..., X_n)$ over all configurations of the input variables that have $X_i = x_i$.

Definition 4. ***Marginal****: The marginal of $f(X_1, X_2, ..., X_n)$ with respect to variable X_i is a function which is denoted $g_{X_i}(x_i)$, and it is obtained by summing over all other variables. More specifically, the marginal with respect to variable X_i at value $x_i \in D_i$ is given by*

$$g_{X_i}(x_i) = \sum_{x_1,...,x_{i-1},x_{i+1},...,x_n} f(X_1, ..., X_i, ..., X_n) \tag{3.8}$$

Following the approach of [61], let $f(X_1, X_2, ..., X_n)$ factors into a product of several local functions, each having some subset of $\{X_1, X_2, ..., X_n\}$ as arguments, specifically, it is assumed that $(X_1, X_2, ..., X_n)$ can be factorized into K factors, namely,

$$f(X_1, X_2, ..., X_n) = \prod_{k=1}^{K} f_k(S_k) \tag{3.9}$$

where $S_k \subseteq \{X_1, X_2, ..., X_n\}$ is the subset of variables associated with the real-valued local factor f_k, i.e. its configuration space. Such factorization is not unique. Function $f(X_1, X_2, ..., X_n)$ itself is a trivial factorization, since it consist of 1 factor.

3.2.2.3 Introduction to Factor Graph Representation

Factor graphs (FGs) provide an efficient way to compute the marginals of a factorizable function using the sum-product algorithm (SPA) [61]. Factor graphs are *bipartite* graphs that represent the factorization of a global function to smaller local functions, e.g. as in equation 3.9. More formally, we provide the definition below as shown in [62]:

Definition 5. ***Factor graph****: Let $f(X_1, X_2, ..., X_n)$ be a decomposable function with K factors, namely $f(X_1, X_2, ..., X_n) = \prod_{k=1}^{K} f_k(S_k)$. The factor graph $G(V, E)$ corresponding to global function f is a bipartite graph, where for every variable X_i, there is a variable node denoted with a circle, and for every factor f_j, there is a factor node denoted with a square. Furthermore, if variable X_i is in the domain of factor f_j an edge is created among them, namely $e_{ij} = (X_i, f_j)$. It is more convenient to write $X_i \in N_G(f_j)$ or equivalently $f_j \in N_G(X_i)$ to denote that variable X_i is argument of factor f_j or in "graph" words, variable node X_i is adjacent with factor node f_j. S_k stands for the subset of the variables of global function f associated with local function f_k.*

For every factorization of function $f(X_1, X_2, ..., X_n)$ there is a unique factor graph $G(V, E)$ and vice-versa, since the mapping between factorizations and factor graphs is one-to-one [61]. Since factor graphs are graphs, they may have cycles or they may have a tree structure. This plays significant role for the convergence of the inference algorithms built on top of these graphs.

Figure 3.6 shows different factor graph examples. Figure 3.6(a) is an example on a directed acyclic factor graph. The corresponding function f for this figure is

$$f(X_1, X_2, X_3, X_4, X_5, X_6) = f_1(X_2, X_5) f_2(X_1, X_3, X_6) f_3(X_1, X_4).$$

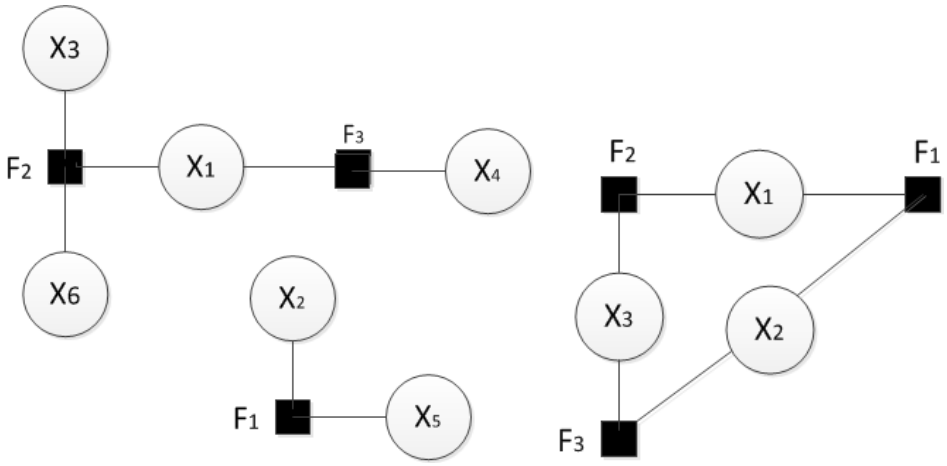
In contrast, figure 3.6(b) illustrates an example on cyclic factor graphs. The corresponding function for this figure is

$$f(X_1, X_2, X_3) = f_1(X_1, X_2) f_2(X_1, X_3) f_3(X_2, X_3).$$

One of our contributions in this thesis is exploiting Bayesian inference through factor graphs to infer confidence scores for human assessments in an efficient way. More details about the proposed inference model is presented in section 3.8.1.

3.2.2.4 Infer.NET Framework

Infer.NET [63] is a framework proposed by Microsoft Research Cambridge for running Bayesian inference in graphical models. Infer.NET provides the state-of-the-art message-



(a) A disconnected, acyclic factor graph example (b) A cyclic factor graph example

Figure 3.6: Different examples on the factor graph representation

passing algorithms and statistical routines needed to perform inference for a wide variety of applications. Infer.NET differs from existing inference software in a number of ways:

- **Rich modeling language**: Support for uni-variate and multivariate variables, both continuous and discrete. Models can be constructed from a broad range of factors including arithmetic operations, linear algebra, range and positive constraints, Boolean operators, Dirichlet-Discrete, Gaussian, and many others. Support for hierarchical mixtures with heterogeneous components.

- **Multiple inference algorithms**: Built-in algorithms include Expectation Propagation, Belief Propagation (a special case of Expectation Propagation), Variational Message Passing and Gibbs sampling.

- **Designed for large scale inference**: In most existing inference programs, inference is performed inside the program - the overhead of running the program slows down the inference. Instead, Infer.NET compiles models into inference source code which can be executed independently with no overhead. It can also be integrated directly into your application. In addition, the source code can be viewed, stepped through, profiled or modified as needed, using standard development tools.

- **User-extendable**: Probability distributions, factors, message operations and inference algorithms can all be added by the user. Infer.NET uses a plug-in architecture which makes it open-ended and adaptable. Whilst the built-in libraries support a wide range of models and inference operations, there will always be special cases where a new

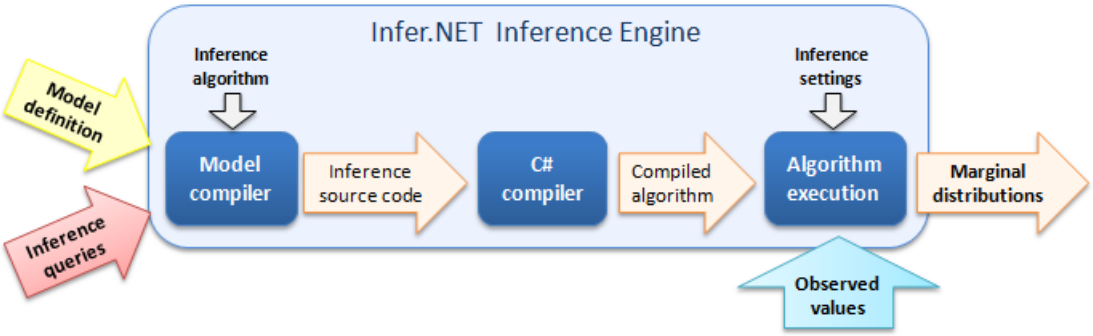


Figure 3.7: Inference process through Infer.NET

factor or distribution type or algorithm is needed. In this case, custom code can be written and freely mixed with the built-in functionality, minimizing the amount of extra work that is needed.

In this thesis, we use Infer.NET to automatically run the inference model in section 3.8.1. Infer.NET works by compiling a model definition into the source code needed to compute a set of inference queries on the model. Figure 3.7 summarizes the inference process.

3.3 Proposed System Architecture

In this section, an overview of the proposed system architecture is presented while providing brief details about its components. Each component is considered separately with more details in following sections. Figure 3.8 shows the architecture of our system. Moreover, figures 3.9 and 3.10 show the sequence of different phases of the proposed approach.

Alignment Module uses GIZA++ [57] word aligner in the offline phase to create an index of source-to-target and target-to-source word alignments from source and target sentences that are stored in different parallel corpora [64, 65, 1]. These alignment models will be used to build phrase alignment on different granularity levels to provide scoring features later.

Features Scoring Module encodes all possible contiguous phrases from the source and target sentences in the translation input. Matches for these phrases are then retrieved from the parallel corpus. For each source(target) match, phrase-alignment matrices are built, based on the generated word alignment model. These matrices are then used to generate data-driven

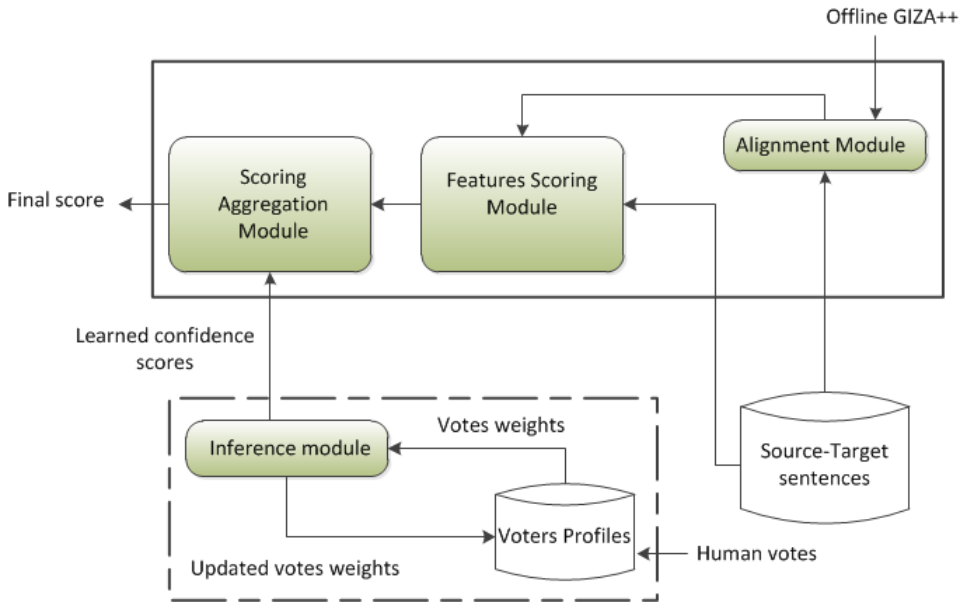


Figure 3.8: Proposed System Architecture. Main modules are rounded with the solid line, and optional modules are rounded with the doted line.

and linguistic scoring features. Having source-to-target and target-to-source features set in the proposed system is the key difference between generating features for translation or for evaluation.

Inference Module runs during the offline phase on the humanly judged translations in the parallel corpus that is used later to extract features. This module operates only if human assessments are available. Each translation can be judged by more than one voter, and the voter can judge many translations. For each translation, there maybe some extreme assessments that should be discarded. Moreover, we could have more judges with various judgments later. Thus, this module captures the effect of judgment variations and converts these variations into judgment confidence scores that will be used later to weigh the proposed features in the *Features Scoring Module*.

Scoring Aggregation Module combines the scores obtained from the *Features Scoring Module* with available weights form human assessments. This module provides a model for aggregating features scores together and finding the set of optimal weights of these features. Human assessments weights have two types: weights from normal human evaluation on a 1-to-5 scale, and weights based on the inferred confidence scores from the *Inference Module*.

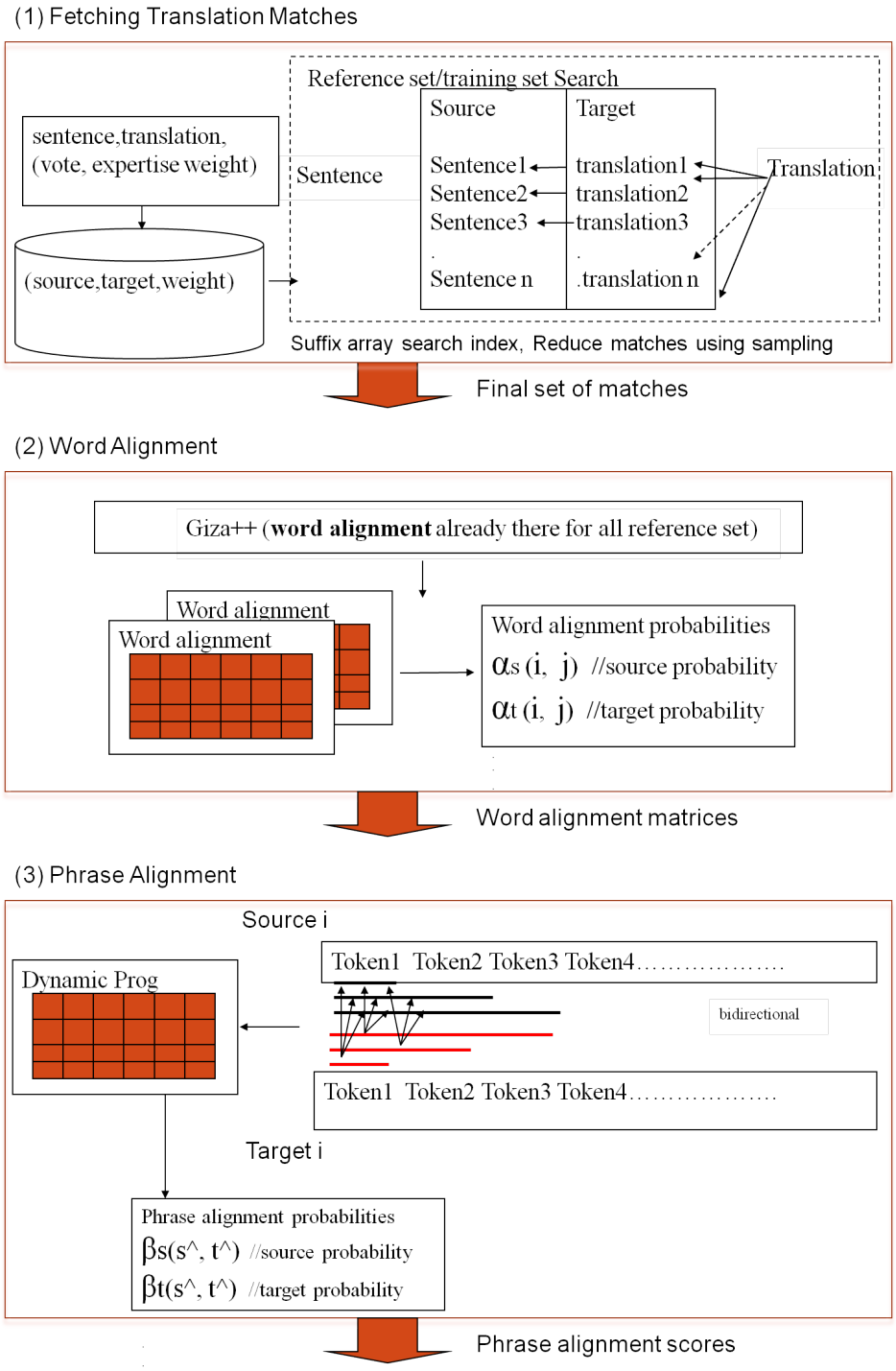


Figure 3.9: Sequence diagram for the alignment operations of the proposed system.

Figure 3.10: Sequence diagram for aggregating scores in the proposed system.

3.4 Fetching Translation Matches

The proposed linguistic and data-driven features to be described in section 3.6 are built based on translation instances that are similar to the input source and target translation. Thus, we need to fetch translation examples from the parallel corpus that approximately match the input source S and target T sentences. We choose to construct a suffix array for indexing each sequence of tokens in the parallel corpus due to its efficiency in storage and search running time as in [19].

Given a translation to be evaluated, the corpus is searched for all partial matches of the translation input. For each sequence of tokens in the source S and target T sentences, we query their respective suffix indexes where retrieved matches should have common tokens with inputs. In addition, the proposed system is capable of locating example instances that are not exact matches of the input but have related semantic to source S and target T sentences. However, locating all matches is not useful for all times. In practice, we simply sample the retrieved matches uniformly. If matches in the corpus are fewer than the desired sample size, then we select all of them. For our experiments, we extract about a few hundred aligned matches to apply features extraction and scoring.

3.5 Alignment Module

After matches are found on the source and target corpora, we determine possible source-target phrase alignments. To alleviate the complexity at run-time, we perform word alignment once during the offline phase and store aligned words as part of the indexed corpus. At run-time, a higher-level alignment between phrases is performed using these word alignments. We observed that we can gain more useful alignment relations if we work on phrases with smaller n-gram lengths as shown in the experimental evaluation (Section 4.4.1).

For each match in the corpus, we load the word alignment matrix for the complete sentence in which the match resides. The alignment matrix A contains scores for all possible word correspondences in this sentence-pair. Each element in A maintains two scores: α_s and α_t where s is a word in the source sentence S and t is a word in the target sentence T. When the word alignments are generated using GIZA++, $p(s_i|t_j)$ will be stored as $\alpha_s(i,j)$. Similarly we store $p(t_j|s_j)$ as $\alpha_t(i,j)$.

The phrase alignment process for each match is composed of two steps.

- Extracting phrases that are consistent with word alignments as indicated in section 3.2.1.3. To find all possible phrases in an efficient way, we used a *backtracking* algorithm that incrementally builds candidates to the solutions, and abandons each partial candidate as soon as this candidate is discovered to be invalid solution.

- Calculating the phrase alignment scores for obtained phrases from the first step. Given a phrase pair $(\hat{s},\hat{t})$, we build the phrase alignment score on top of the word alignment scores as follows:

$$\beta_s(\hat{s},\hat{t}) = -\sum_{s_i \in \hat{s}} \sum_{t_j \in \hat{t}} \ln \alpha_s(i,j) \tag{3.10}$$

$$\beta_t(\hat{s},\hat{t}) = -\sum_{s_i \in \hat{s}} \sum_{t_j \in \hat{t}} \ln \alpha_t(i,j) \tag{3.11}$$

  In the following section, we use $\beta_s(\hat{s},\hat{t})$ and $\beta_t(\hat{s},\hat{t})$ to define feature functions used for evaluating translations.

3.6 Features Scoring Module

In this section, we present the proposed set of feature functions that are used to score translations. These features can be grouped into three classes; Alignment features, coverage features and statistics features. For the *Alignment features*, we exploit the phrase alignment output from the alignment module (Sections 3.6.1 and 3.6.2). *Coverage features*, measure to what extent the input translation is covered by matched phrases from the parallel corpus (Section 3.6.3). *Statistics features* is the third class which has features that are built on statistics from the parallel corpus (Section 3.6.4).

3.6.1 Outside Source/Target Alignment Features

Given a pair of source and target sentences, each source phrase is assumed to be aligned to every possible target phrase with a certain probability. This alignment not only implies that words within the source phrase are aligned to words within the target phrase, but also

that the remainder of the source sentence not specified by the source phrase is aligned to the remainder of the target sentence not specified by the target phrase.

Let us assume the set of phrases in the source S and target T sentences that are outside the phrase alignment between $\hat{s}$ and $\hat{t}$ are $\hat{s}_{out}$ and $\hat{t}_{out}$ respectively. The following equations define outside source and target alignment feature functions.

$$f_{OS}(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}}) = -\ln \sum_{\hat{s}_i \in \hat{s}_{out},\acute{\hat{s}},\acute{\hat{t}}} \acute{\phi}_{OS}(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{s}}) = -\ln \sum_{\hat{s}_i \in \hat{s}_{out},\acute{\hat{s}},\acute{\hat{t}}} \frac{\sum_{\hat{t}_j \in \hat{t}_{out},\acute{\hat{s}},\acute{\hat{t}}} \beta_{\hat{t}}(\hat{s}_i,\hat{t}_j)}{\sum_{\hat{t}_j,\acute{\hat{s}},\acute{\hat{t}}} \beta_{\hat{t}}(\hat{s}_i,\hat{t}_j)} \tag{3.12}$$

Similarly,

$$f_{OT}(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}}) = -\ln \sum_{\hat{t}_j \in \hat{t}_{out},\acute{\hat{s}},\acute{\hat{t}}} \acute{\phi}_{OT}(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}}) = -\ln \sum_{\hat{t}_j \in \hat{t}_{out},\acute{\hat{s}},\acute{\hat{t}}} \frac{\sum_{\hat{s}_i \in \hat{s}_{out},\acute{\hat{s}},\acute{\hat{t}}} \beta_{\hat{s}}(\hat{s}_i,\hat{t}_j)}{\sum_{\hat{s}_i,\acute{\hat{s}},\acute{\hat{t}}} \beta_{\hat{s}}(\hat{s}_i,\hat{t}_j)} \tag{3.13}$$

The main purpose of these functions is to show the tendency of phrase alignment to be concentrated outside the sentence boundaries.

3.6.2 Inside Source/Target Alignment Features

Analogously, we can show the tendency of source/target phrase alignment to be concentrated inside the sentence boundaries using similar formulas to equations 3.12 and 3.13. Let us assume the set of phrases in the source and target sentences that are inside the phrase alignment are $\hat{s}_{in}$ and $\hat{t}_{in}$ respectively. The following equations define inside source and target alignment feature functions.

$$f_{IS}(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}}) = -\ln \sum_{\hat{s}_i \in \hat{s}_{in},\acute{\hat{s}},\acute{\hat{t}}} \acute{\phi}_{IS}(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}}) = -\ln \sum_{\hat{s}_i \in \hat{s}_{in},\acute{\hat{s}},\acute{\hat{t}}} \frac{\sum_{\hat{t}_j \in \hat{t}_{in},\acute{\hat{s}},\acute{\hat{t}}} \beta_{\hat{t}}(\hat{s}_i,\hat{t}_j)}{\sum_{\hat{t}_j,\acute{\hat{s}},\acute{\hat{t}}} \beta_{\hat{t}}(\hat{s}_i,\hat{t}_j)} \tag{3.14}$$

$$f_{IT}(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}}) = -\ln \sum_{\hat{t}_j \in \hat{t}_{in},\acute{\hat{s}},\acute{\hat{t}}} \acute{\phi}_{IT}(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}}) = -\ln \sum_{\hat{t}_j \in \hat{t}_{in},\acute{\hat{s}},\acute{\hat{t}}} \frac{\sum_{\hat{s}_i \in \hat{s}_{in},\acute{\hat{s}},\acute{\hat{t}}} \beta_{\hat{s}}(\hat{s}_i,\hat{t}_j)}{\sum_{\hat{s}_i,\acute{\hat{s}},\acute{\hat{t}}} \beta_{\hat{s}}(\hat{s}_i,\hat{t}_j)} \tag{3.15}$$

3.6.3 Source/Target Coverage Features

Let $\hat{s}$ and $\hat{t}$ represent the source and target phrases for the input translation where S and T represent the entire input source and target sentences. The ratios of $\frac{|\hat{s}|_{words}}{|S|_{words}}$ and $\frac{|\hat{t}|_{words}}{|T|_{words}}$ represent the translation coverage by each phrase alignment (1 means covering the entire sentence where 0 means no coverage at all). Formally, we define this feature function for source and target phrases, respectively, as follows:

$$f_{sc}(\hat{s},\hat{t}) = \phi_{sc}(\hat{s},\hat{t}) = \frac{|\hat{s}|_{words}}{|S|_{words}} \tag{3.16}$$

and

$$f_{tc}(\hat{s},\hat{t}) = \phi_{tc}(\hat{s},\hat{t}) = \frac{|\hat{t}|_{words}}{|T|_{words}} \tag{3.17}$$

3.6.4 Source/Target Statistics Features

We propose here to use correlation between frequencies of the source and target phrases in the corpus as a feature. Let $C(\hat{s})$ be the number of occurrences of phrase $\hat{s}$ in the corpus. Similarly, let $C(\hat{t})$ be the number of occurrences of phrase $\hat{t}$ in the corpus.

$$f_{freq}(\hat{s},\hat{t}) = \phi_f(\hat{s},\hat{t}) = \frac{(C(\hat{s}) - C(\hat{t}))^2}{(C(\hat{s}) + C(\hat{t}) + 1)^2} \tag{3.18}$$

This formula needs only, for each source-target aligned instance, to store the count of sentences in the training corpus in which this instance appeared.

In order to offset the tendency to prefer short translations, we balance the overall evaluation score by including a feature that simply counts the number of words in both source $\hat{s}$ and target $\hat{t}$ phrases. This feature uses correlation between the number of words in $\hat{s}$ and $\hat{t}$ as follows:

$$f_{WR}(s_i,t_j) = \phi_{WR}(s_i,t_j) = \frac{(|\hat{s}|_{words} - |\hat{t}|_{words})^2}{(|\hat{s}|_{words} + |\hat{t}|_{words} + 1)^2} \tag{3.19}$$

Similarly for the number of characters in $\hat{s}$ and $\hat{t}$, we have:

$$f_{CR}(s_i,t_j) = \phi_{CR}(s_i,t_j) = \frac{(|\hat{s}|_{char} - |\hat{t}|_{char})^2}{(|\hat{s}|_{char} + |\hat{t}|_{char} + 1)^2} \tag{3.20}$$

3.7 Scoring Aggregation Module

Given the phrase pair $(\hat{s},\hat{t})$ and an aligned match $(\acute{\hat{s}},\acute{\hat{t}})$, we define $\phi(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}})$ to provide a unified representation for feature functions that are described in section 3.6 where

$$\phi(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}}) = \begin{cases} f(\hat{s},\hat{t}) & \text{if } (\acute{\hat{s}},\acute{\hat{t}}) \text{ doesn't exist} \\ f(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}}) & \text{if } (\acute{\hat{s}},\acute{\hat{t}}) \text{ exists} \end{cases} \tag{3.21}$$

In general, some features may have higher impact than others. Formally, we can capture this by introducing a weight θ_l for each feature function $\phi_l(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}})$ that let us scale the contribution of this feature. Reducing the set of features to one log-linear model considerably simplifies estimating these θ parameters during optimization using log-likelihood algorithm. However, the proposed approach can't fit the log-linear model. Fortunately, [19] demonstrated an approximation method for the log-linear model to score translation units for example-based machine translation. We customized this method to define the overall scoring function $E(\hat{s},\hat{t},\theta)$ for the phrase pair $(\hat{s},\hat{t})$ as follows:

$$E(\hat{s},\hat{t},\theta) = \ln \sum_{\acute{\hat{s}},\acute{\hat{t}}} e^{\sum_{l,\hat{s},\hat{t}} \theta_l \acute{\phi}_l(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}})} \tag{3.22}$$

To find the optimal value of θ parameters, the derivative of equation 3.22 is needed. We approximated the derivative of this equation using first-order Taylor series as shown in [19]. Based on that, the derivative of the model with respect to θ is:

$$E(\hat{s},\hat{t},\acute{\theta}) \approx E(\hat{s},\hat{t},\theta) + \sum_q (\acute{\theta}_q - \theta_q)\frac{\partial}{\partial \theta_q} E(\hat{s},\hat{t},\theta) \approx \ln \sum_{\acute{\hat{s}},\acute{\hat{t}}} e^{\sum_{l,\hat{s},\hat{t}} \theta_l \acute{\phi}_l(s_i,\acute{\hat{s}},\acute{\hat{t}},\acute{\hat{t}})} + \sum_q (\acute{\theta}_q - \theta_q) \left\{ \frac{\sum_{\acute{\hat{s}},\acute{\hat{t}}} \acute{\phi}_q(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}}) e^{\sum_{l,\hat{s},\hat{t}} \theta_l \acute{\phi}_l(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}})}}{\sum_{\acute{\hat{s}},\acute{\hat{t}}} e^{\sum_{l,\hat{s},\hat{t}} \theta_l \acute{\phi}_l(\hat{s},\acute{\hat{s}},\hat{t},\acute{\hat{t}})}} \right\} \tag{3.23}$$

The optimal values of θ can be estimated using any numerical analysis method. After that, the final evaluation score e_f can be easily assembled, given the source phrases $\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_n$ and target phrases $\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_m$, as follows:

$$e_f = \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_{ij} E(\hat{s}_i, \hat{t}_j, \acute{\theta}) \tag{3.24}$$

where λ_{ij} represents the human assessment weight for the phrase pair $(\hat{s}_i, \hat{t}_j)$. This weight plays an important role only if these assessments are available. Formally, we can define λ_{ij} as follows:

$$\lambda_{ij} = \begin{cases} 0.5 + \lambda_{ij}^{normal} & \text{available human assessments without applying inference} \\ 0.5 + \lambda_{ij}^{infer} & \text{available human assessments with applying inference} \\ 1 & \text{no human assessments} \end{cases} \tag{3.25}$$

λ_{ij}^{normal} is a score provided by judges on the 1-to-5 scale. Each score is divided by the maximum value (e.g. 5 in our case) to be normalized in the range $[0,1]$. Alternatively, a confidence score λ_{ij}^{infer} can be obtained by applying the inference model on available human assessments as will be shown in section 3.8.2. These confidence scores are probabilities with values in the range $[0,1]$, so no need for normalization.

3.8 Inference Module

Human evaluation is usually performed by language experts and native speakers. However, these human assessments are expensive to collect. In addition, sometimes different human assessments disagree for the same sentence.

Here comes the need for a model that learns the uncertainties in relevance between human assessments based on these discrepancies. In designing the inference model, we start by thinking about the nature of the human assessments. Given a translation, an assessment can be categorized as one of ordinal scores given by a voter from the set $\{1,2,...,L\}$ where 1 represents a bad translation and L represents an excellent translation.

Each voter provides only one score per translation, and each translation can be evaluated by many voters. These votes are stored in the corresponding voters profiles to be used by the inference module (see figure 3.8). It is assumed that all scores are equally treated, with no bias, which means that we are confident of the voters experience (an assumption that can be relaxed later in the model). Due to the scarcity nature of human scores, each human score can be considered to provide a bit of evidence about the quality of the translation. The more human scores we have, the more we get confident about the translation score.

In the following subsections, we provide the details of the proposed inference model and its factor graph to estimate the confidence of human scores in an efficient way.

3.8.1 Probabilistic Inference Model

We present a probabilistic model to infer confidence scores from human assessments. This model can be trained from previously judged translations where both voters and translations are defined by random variables to be easily represented with factor graphs.

3.8.1.1 Scoring Model

Initially, let us assume that the inference model receives tuples in the form (x,y,r). x, y and r are defined as follows assuming that R is the set of real numbers.

- $x \in R$ represents the voter confidence in general.
- $y \in R$ represents the translation score in general.
- $r \in R$ represents the score given by the voter to the translation.

Similar to [66], we define a voter confidence variable as

$$s = ux \tag{3.26}$$

where u is a latent voter confidence descriptor. Similarly, we define the translation scoring variable $t = vy$ where v is a latent translation scoring coefficient. Now the probability of score r is modeled as

$$p(r|s,t) = N(r|s,t,\beta^2) \tag{3.27}$$

where N is a normal distribution, and β is the standard deviation of the observation noise. Thus, we adopt a form in which the expected score given by a voter to a certain translation is given by the inner product of the voter and translation variables.

3.8.1.2 Prior Estimation

The model parameters to be learned are the variables u and v which determine how voters and translations are mapped to the random variable space. We represent our prior beliefs about the values of these parameters by independent Gaussian distributions, i.e, $p(u) = N(\mu_u, \sigma_u^2)$ and $p(v) = N(\mu_v, \sigma_v^2)$.

Each prior distribution represents a factor in the factor graph. We choose this prior distribution because it reduces memory requirements to two parameters (a mean and standard deviation) and it allows us to perform efficient inference as will be shown.

3.8.1.3 Adaptation to Ordinal Scores

A common scenario is that voters provide feedback about translations they like or dislike via an ordinal score. These scores can only be compared, but not subtracted from one another. In addition, each voter's interpretation of the scale may be different and the mapping from score to latent scoring may not be linear.

We assume that for each voter-translation pair for which data is available, we observe a score $l \in \{1,2,...,L\}$. We relate the latent scoring variable r to the score l via a cumulative threshold model [67]. For each voter u, we maintain voter-specific thresholds $b_u \in R^{L-1}$ which divide the latent scoring axis into L consecutive intervals $(b_{u(i-1)}, b_{u(i)})$ of varying length each of which representing the region in which this voter gives the same score to a translation. Formally, we define a generative model of a score as

$$p(l=a|b_u,r)=\begin{cases}\prod_{i=1}^{a-1} I(r>b^{\sim}_{u(i-1)})\prod_{i=a}^{L-1} I(r<b^{\sim}_{u(i-1)}) & \text{if } 1<a<L\\ \prod_{i=1}^{L-1} I(r<b^{\sim}_{u(i-1)}) & \text{if } a=1\\ \prod_{i=1}^{L-1} I(r>b^{\sim}_{u(i-1)}), & \text{if } a=L\end{cases} \tag{3.28}$$

where $p(b^{\sim}_{u(i)}|b_{u(i)})=N(b^{\sim}_{u(i)};b_{u(i)},\tau^2)$ and we place an independent Gaussian prior on the thresholds $p(b_{u(i)})=N(b_{u(i)};\mu_{u(i)},\sigma^2)$. The indicator function $I(.)$ is equal to 1 if the proposition in the argument is true and 0 if it is false. Inferring these thresholds for each voter allows us to discard extreme or inconsistent scores compared to the expected range of scores of her.

3.8.2 Factor Graph Representation

Given a stream of scoring tuples (x,y,r), we train the model in order to learn posterior distributions over the values of the parameters u and v. This can be accomplished efficiently by message passing [63]. The model described in section 3.8.1.1 can be further factorized by introducing some intermediate latent variables z_k to represent the result of the inner product of s_k and t_k where k represents a certain voter. That is,

$$p(z_k|s_k,t_k)=I(z_k=s_kt_k) \tag{3.29}$$

Now the latent score over a set of human scores is given by

$$p(r^{\sim}|z,b)=I(r^{\sim}=\sum_k z_k+b) \tag{3.30}$$

From the scoring model in section 3.8.1.1, we can estimate $p(s_k|u_k,x_k)$ as $I(s_k=u_kx_k)$ and $p(t_k|v_k,y_k)$ as $I(t_k=v_ky_k)$. Therefore, the joint distribution of all the variables can be factorized as

$$p(s,t,u,v,z,r^{\sim},r|x,y)=p(r|r^{\sim})p(r^{\sim}|z,b)p(u)p(v)\prod_{i=1}^{k}p(z_k|s_k,t_k)p(s_k|u_k,x_k)p(t_k|v_k,y_k) \tag{3.31}$$

Figure 3.11 shows the factor graph of the inference model. Inferred $r^{\sim}$ values are used to

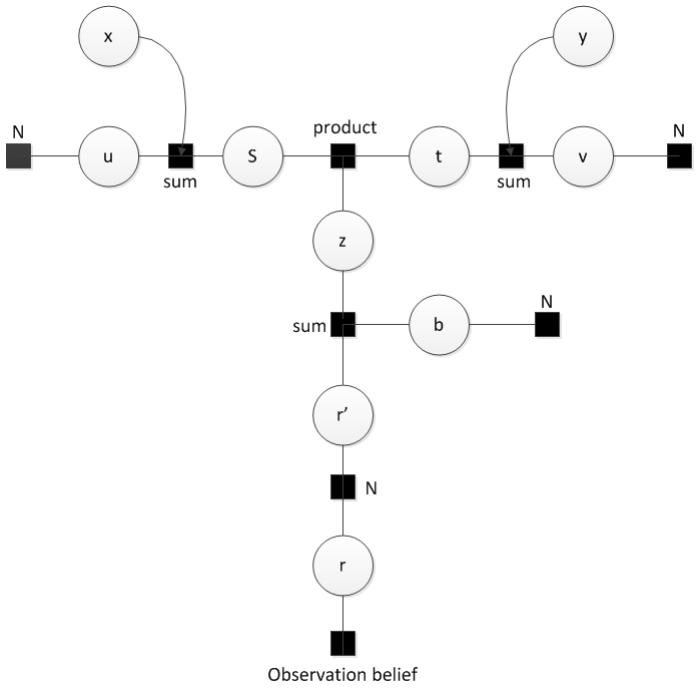


Figure 3.11: Factor graph of the inference model

provide *confidence* weights λ_{ij}^{infer} for the aggregated features score (Section 3.7). Moreover, voters profiles are updated with these confidence scores to be used for further evaluation.

3.9 Complexity Analysis

Given a translation pair (S,T), let n be the number of possible phrases in the source sentence S and m be the number of possible phrases in the target sentence T. Let C be the number of sentences in the parallel corpus and l be the maximum number of words inside each phrase. First, the time required to retrieve translation matches for (S,T) is $O(\log C)$ [19] because we construct a suffix index for the whole parallel corpus in the offline mode.

Given a certain phrase pair $(\hat{s},\hat{t})$, the time required to build a phrase alignment matrix online is $O(l^2 * \log C)$ because we search in the $\log C$ matches for each pair of aligned words in $(\hat{s},\hat{t})$. Therefore, the time required to build phrase alignment matrices for all possible phrase pairs in the input translation is $O(n*m*l^2*\log C)$. However, the values of n, m, and l are much smaller than $\log C$. Thus, the online time complexity is reduced to $O(\log C)$.

If human assessments are available, we apply the inference model on these assessments in the offline mode. The time required to apply the inference model on each match in the

$\log C$ matches is *constant* [66]. Thus, the time required to apply the inference model for all matches is $O(constant * \log C) = O(\log C)$ which is the offline time complexity.

3.10 Conclusion

In this chapter, the proposed approach for machine translation evaluation was presented. First, the theoretical foundations upon which the proposed approach is built are presented briefly. These foundations include the theoretical bases of alignment models and Bayesian inference techniques. Then, a brief overview of the system architecture was shown. After that, we provided the details of the work which presents a novel evaluation approach for machine translation outputs. The core approach draws from proposing a set of linguistic and data-driven features that can be weighted by human assessments if these assessments are available. The proposed approach provides an inference model to infer credible human scores to be used as weights. Finally, the complexity analysis of the proposed approach is discussed.

In the next chapter, the proposed approach is evaluated using experiments on standard datasets.

CHAPTER 4

EXPERIMENTAL EVALUATION

In this chapter, we present and discuss the evaluation procedure and the results of the proposed approach. First, we present the metrics used throughout the evaluation in section 4.1. Implementation tools and configurations are presented in section 4.2. Then, details about the used datasets and the evaluation procedure are provided in section 4.3. Finally, the evaluation results are presented in detail in section 4.4.

4.1 Evaluation Metrics

The metrics we used to evaluate the output of the proposed machine translation evaluation approach can be divided into two main categories: accuracy metrics and time metrics. The accuracy metrics include two different metrics: correlation with human judgment and cumulative distribution function (CDF) of correlation.

In traditional automatic metrics, the authors used only correlation with human judgment. However, we believe that CDF of correlation is necessary to have a good comprehensive comparison between different combinations of features.

4.1.1 Correlation with Human Judgment

Several metrics can be used for measuring correlation with human judgments. The main metric used is Spearman's score correlation coefficient. We opted for Spearman rather than Pearson because it makes fewer assumptions about the data. Importantly, it can be applied to ordinal data. In general, Spearman correlation coefficient is equivalent to Pearson correlation

in scoring. We use the simplified Spearman form in [64] as follows:

$$\rho = 1 - \frac{6\sum d_{ij}^2}{n(n^2-1)} \tag{4.1}$$

where d_{ij} is the euclidean distance and can be substituted by the difference between the score r_i given by the evaluation metric and the human judgment score r_j for a certain translation, and n is the number of translations. The possible values of ρ range between 1 (where all translations are scored in the same manual order) and -1 (where the translations are scored in the reverse order). The higher correlation we have, the closer to human evaluation we are.

4.1.2 Cumulative Distribution Function (CDF) of Correlation

As observed, the correlation metrics are concerned only with the intersection between the human judgment and the automated evaluation output. To evaluate the approach better, a metric that considers the overall picture should be used such as Cumulative Distribution Function.

CDF here describes the probability that a Spearman correlation ρ with a given probability distribution will be found to have a value less than or equal to t_ρ where t_ρ is used to predict the quality of estimation.

4.1.3 Response Time

To study the usability of the introduced approach, its execution time is studied. According to the context, total response time is used. This means all the time needed to evaluate a machine translation output. This metric is used when evaluating the proposed approach using different configurations of parameters. Also, it provides a useful indication about the applicability of the approach compared to other traditional evaluation metrics.

4.2 Tools and Implementation

The proposed approach is implemented using Java 6.0 as the programming platform. For word alignment, we trained IBM Model-4 [17] using GIZA++ [57]. Using the monolingual corpus, we trained many English language models with different N-grams as specified in [19]. For the online operation, our approach uses extracted phrases corresponding to both tokenized monolingual corpus and word alignments from the offline phase.

The confidence of human scores are inferred using the Expectation-Propagation and Variational Message Passing algorithms as in [60] to weigh initial features scores. We use Infer.NET [63] framework to implement the inference module because it provides a variety of built-in Bayesian inference packages.

The required storage for word-to-word alignment process via GIZA++ [57] is exponential in the size of training corpus. Moreover, it is the dominating task during the offline preparation and needs high memory consumption. Thus, once alignment is done we index output files to be used for features extraction later. The whole pre-processing bottleneck is handled on the server side only. Whenever inferred confidence of new human scores is detected, the server updates the users' profile.

The evaluation was conducted on a Dell Laptop with a core-i7 processor and 8GB RAM. After running the considered approach on the different datasets, CDFs and statistical measures of accuracy and time metrics are provided.

4.3 Datasets and Human Judgments

We use two types of datasets: Europarl [65] multilingual corpus as a large dataset, and WMT [64] corpora as small datsets. The majority of the training data in the two datasets was drawn from the Europarl corpus. Additional training data was taken from the News Commentary corpus. The News Commentary test set differs from the Europarl data in various ways. The domain is general politics, economics and science. However, it is also mostly political content (even if not focused on the internal workings of the European Union) and opinion. The test data was drawn from a segment of the Europarl corpus which is excluded from the training data.

Language	Sentences	Words
Danish (Da)	1,032,764	27,153,424
German (De)	1,023,115	27,302,541
Greek (El)	746,834	27,772,533
English (En)	1,011,476	28,521,967
Spanish (Es)	1,029,155	30,007,569
French (Fr)	1,023,523	32,550,260
Finnish (Fi)	941,890	18,841,346
Italian (It)	979,543	28,786,724
Dutch (Nl)	1,042,482	28,763,729
Portuguese (Pt)	1,014,128	29,213,348
Swedish (Sv)	947,493	23,535,265

Table 4.1: The number of sentences and words in Eurporl 2005 for different European languages.

In this section, we give details about the datasets used to evaluate the performance of the proposed approach.

4.3.1 Europarl Dataset

Europarl corpus [65] is a collection of the proceedings of the European Parliament dating back to 1996. Europarl corpus was mainly proposed to aid the research in statistical machine translation field, but since it was made available in its initial release in 2005, it has been used for many other natural language problems: word sense disambiguation, information extraction, etc.

Till 2005, this corpus comprised of about 30 million words for each of the 11 official languages of the European Union: Danish (Da), German (De), Greek (El), English (En), Spanish (Es), Finnish (Fi), French (Fr), Italian (It), Dutch (Nl), Portuguese (Pt), and Swedish (Sv). Europarl has been expanded to include more than 60 million words per language at 2012. In this thesis, Europarl 2005 was used as a large dataset to evaluate the proposed linguistic and data-driven features. We focused on two challenging parallel corpora: French-to-English (Fr-to-En) and Spanish-to-English (Es-to-En).

Table 4.1 summarizes the statistics of sentences and words for different European languages in 2005.

4.3.2 WMT Datasets

WMT workshops [64] usually provide parallel corpora of European language pairs in their annual event. These corpora are offered to evaluate the translation quality of proposed MT systems in shard task between these systems. The shared task data included training, development and testing sets from the Europarl multilingual and the News Commentary data. The parallel sentences in this shared task data were evaluated using both manual evaluation and automatic metrics.

For the manual evaluation, the workload was distributed across a number of people, including participants in the shared task, interested volunteers, and a small number of paid annotators. More than 100 people participated in the manual evaluation, with 75 of those people putting in at least an hour's worth of effort. The main goal of this manual evaluation was to collect data which could be used to assess how well automatic metrics correlate with human judgments.

In this thesis, WMT 2007 and WMT 2013 datasets were used as small datasets to evaluate the proposed approach. It is assumed that submissions from different MT systems in the shared task are coming from different translators where each MT system represents a translator. We believe that WMT datasets are suitable to evaluate the proposed approach for the following reasons:

- It supports a wide range of language pairs (e.g. German-to-English, Spanish-to-English and French-to-English). We focused only on French-to-English (Fr-to-En) and Spanish-to-English (Es-to-En) datasets to be consistent with Europarl datasets in section 4.3.1.

- More than ten automatic evaluation metrics were applied and their correlation with the human scores was calculated. To evaluate the performance of our approach, we compare it with commonly used evaluation metrics in WMT shared tasks, such as BLEU [23], GTM [26], 1-TER [27], and METEOR [22]. METEOR has special importance because it has been shown to correlate better with the human perception of translation quality in previous research work [43].

- Extensive human evaluation was carried out per each translated sentence which allowed high confidence in the given score. Human scores were randomly split into 66% for training the inference model, and the remaining scores were used for calcu-

	French-to-English			Spanish-to-English		
	Sentence	Words	Distinct words	Sentence	Words	Distinct words
Europarl Train	1,288,901	65,791,528	613,005	1,259,914	64,973,029	612,920
News Train	43,194	1,935,265	131,782	51,613	2,339,340	155,058
Europarl Test	2000	53,981	10,186	2000	55,380	10,451
News Test	2007	49,820	11,244	2007	50,771	10,948

Table 4.2: Statistics of French-to-English and Spanish-to-English datasets from WMT 2007.

	French-to-English			Spanish-to-English		
	Sentence	Words	Distinct words	Sentence	Words	Distinct words
Europarl Train	2,007,723	60,125,563	140,915	1,965,734	56,895,229	176,258
News Train	157,168	4,928,135	69,028	174,441	5,116,388	84,273
Europarl Test	2500	120,194	17,230	2500	67,835	17,940
News Test	3000	130,810	19,935	3000	130,810	18,610

Table 4.3: Statistics of French-to-English and Spanish-to-English datasets from WMT 2013.

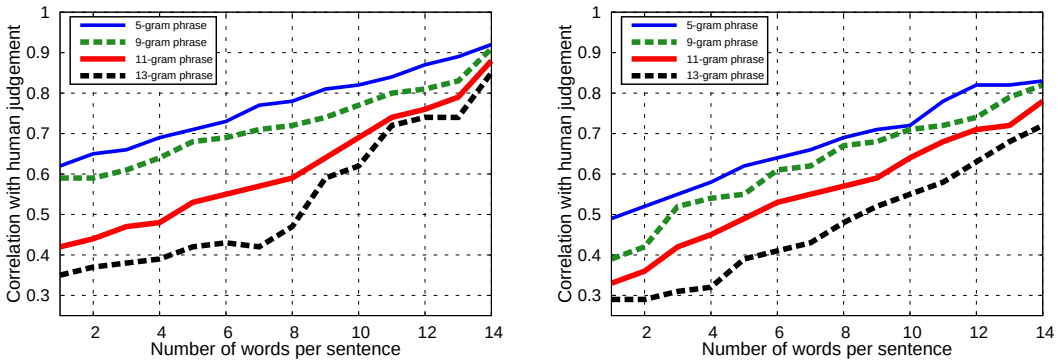
lating the correlation during testing. This process was repeated 5 times to generate different splits, and we calculated the average score.

We mainly used WMT 2007 [64] dataset to perform detailed evaluation. Then, we also used WMT 2013 [45] dataset to compare the proposed approach using its best combination of features to traditional evaluation metrics. Tables 4.2 and 4.3 show some statistics about WMT 2007 and WMT 2013 datasets respectively.

4.4 Performance Evaluation

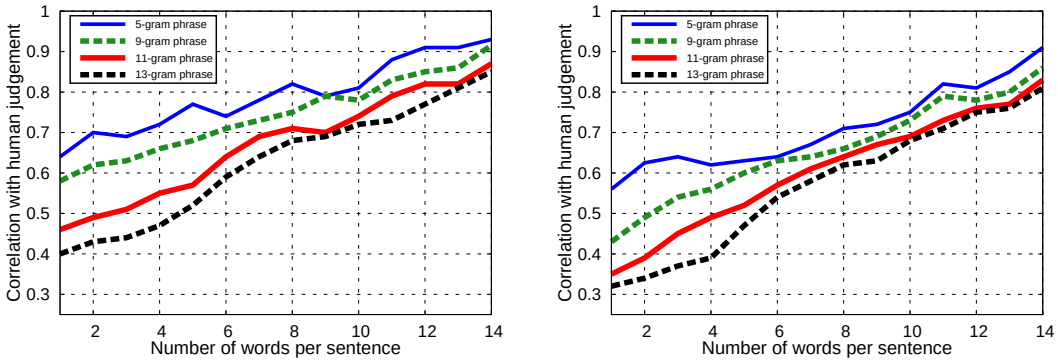
In this section, the results of the evaluation are presented. The effect of different parameters on the performance is studied. Then, a comparison with traditional evaluation metrics and quality estimation approaches is provided.

Evaluation results are divided into two parts. The first part evaluates the performance of the proposed approach using the linguistic and data-driven features and without including human assessments (sections 4.4.1, 4.4.2 and 4.4.3). These features are examined in detail on different corpus scales using WMT and Europarl datasets. The second part studies the effect of incorporating human assessments on the overall accuracy (sections 4.4.4 and 4.4.5). In this part, only WMT datasets are used because no human assessments are available for



(a) French-to-English dataset (b) Spanish-to-English dataset

Figure 4.1: Effect of different N-gram lengths on the correlation with human judgments using WMT 2007 datasets.



(a) French-to-English dataset (b) Spanish-to-English dataset

Figure 4.2: Effect of different N-gram lengths on the correlation with human judgments using Europarl 2005 datasets.

Europarl corpus.

4.4.1 Effect of N-gram Length

N-gram is defined as a contiguous sequence of N items from a given sequence of text. The items can be letters, words or base pairs according to the application. In the proposed approach, N-gram is made out of N consecutive words for each phrase to find the suitable unit for phrase alignment that increases the accuracy.

Figure 4.1 demonstrates the effect of choosing different values of N on the accuracy of the proposed approach using WMT 2007 datasets. It has been found that decreasing the value of N leads to a steady improvement in the correlation with human judgments. Using lower values of N increases the possibility of more matches between source and target corpora

Dataset	WMT 2007	Europarl 2005
French-to-English dataset	0.92	0.93
Spanish-to-English dataset	0.83	0.89

Table 4.4: Correlation with human judgments for WMT 2007 and Europarl 2005 datasets using 5-gram configuration.

Dataset	WMT 2007	Europarl 2005
French-to-English dataset	0.91	0.915
Spanish-to-English dataset	0.82	0.86

Table 4.5: Correlation with human judgments for WMT 2007 and Europarl 2005 datasets using 9-gram configuration.

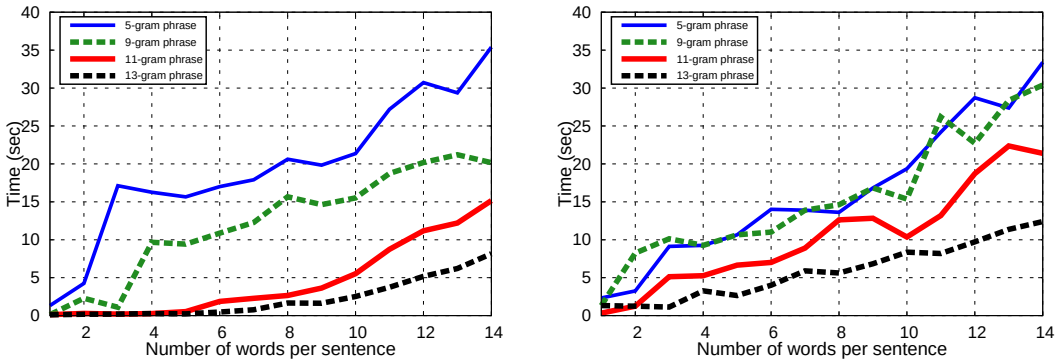
which leads to more accuracy in terms of higher correlation. This finding has been confirmed by results from large Europarl 2005 datasets as shown in figure 4.2.

However, there is a tradeoff between accuracy and response time depending on the N-gram level used. By decreasing the value of N, the system will incur an increase in response time. Figures 4.3 and 4.4 show the effect of choosing different values of N on the response time of the proposed approach using WMT 2007 and Europarl 2005 datasets respectively. If the value of n is chosen to be less than 5, the system response time will increase by around 200%. Therefore, we select 5-gram and 9-gram values as default configurations for the remaining performance studies. Tables 4.4 and 4.5 summarize the accuracy performance of the proposed approach using these configurations. It is be noted that the proposed approach is cost-effective since one of its main applications is to select good enough translations for human post-editing from large-scale batches of translations.

4.4.2 Effect of Different Combinations of Features

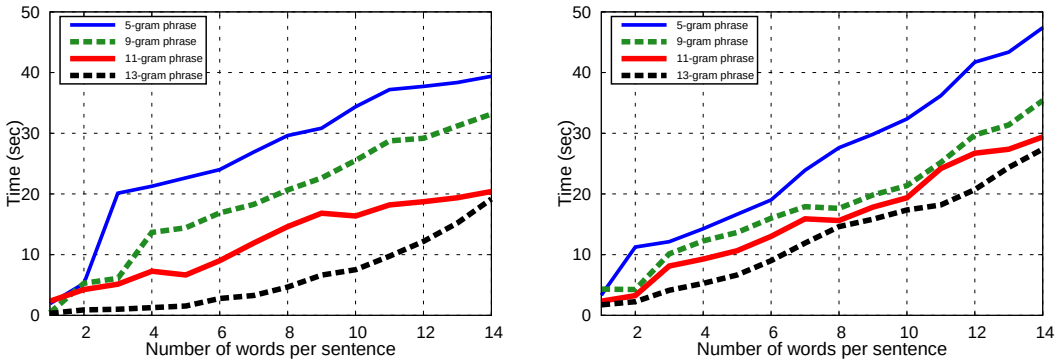
As stated in section 3.6, we proposed a novel set of feature functions for scoring translations. To study the effect of these features, we group them into classes and try different combinations of these classes. We can categorize features into three main classes as follows:

- **Alignment Features (AF)**:
  - Outside source/target alignment probabilities, f_{OS}/f_{OT}.
  - Inside source/target alignment probabilities, f_{IS}/f_{IT}.
- **Coverage Features (CF)**:



(a) French-to-English dataset (b) Spanish-to-English dataset

Figure 4.3: Effect of different N-gram lengths on the response time using WMT 2007 datasets.



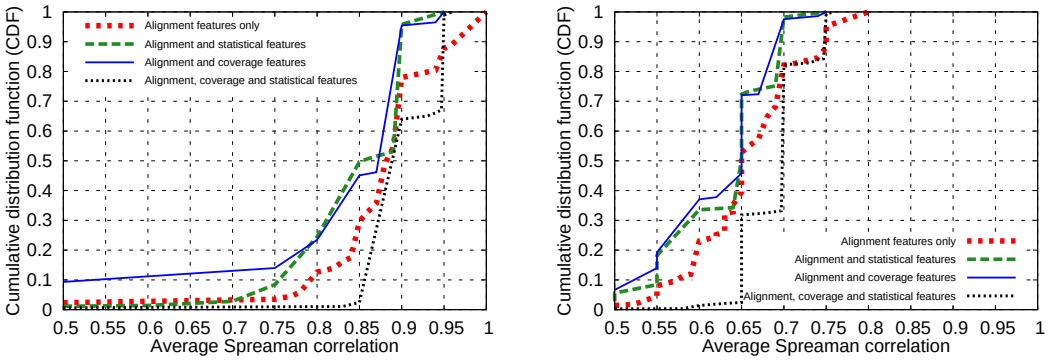
(a) French-to-English dataset (b) Spanish-to-English dataset

Figure 4.4: Effect of different N-gram lengths on the response time using Europarl 2005 datasets.

  - Ratio of the number of source words covered by the target sentence, f_{SC}.
  - Ratio of the number of target words covered by the source sentence, f_{TC}.

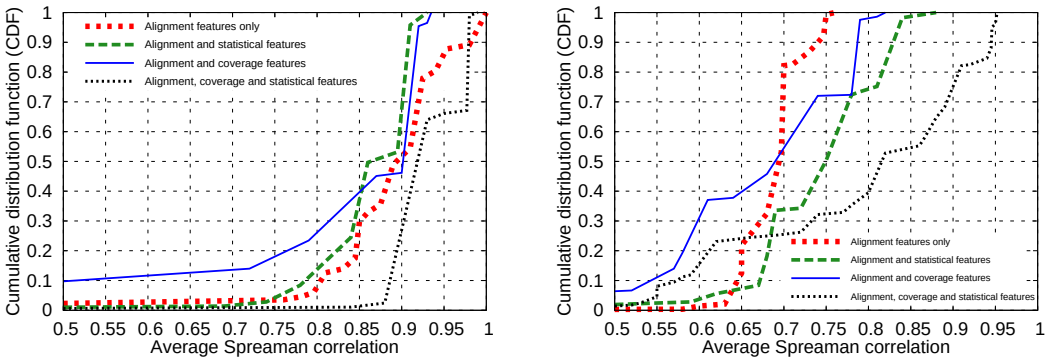
- **Statistical Features (SF)**:
  - Ratio of Source/Target words, f_{WR}.
  - Ratio of Source/Target characters, f_{CR}.
  - Source/target sentence length/occurrences in corpus.
  - Average source/target word length.

Figures 4.5 and 4.6 show the CDF of correlation with human judgment using these different classes of features. We observe that the best accuracy is achieved when all features are included. Moreover, features extraction from large corpora leads to more accuracy. These



(a) French-to-English dataset (b) Spanish-to-English dataset

Figure 4.5: CDF of correlation with human judgments for different combinations of features using WMT 2007 datasets.



(a) French-to-English dataset (b) Spanish-to-English dataset

Figure 4.6: CDF of correlation with human judgments for different combinations of features using Europarl 2005 datasets.

observations are validated by the two language pairs. We can show that probability of obtaining average correlation value lower than 0.65 is almost zero which ensures a robust lower threshold on the accuracy. Tables 4.6 and 4.7 provide a summary for average correlation values using different classes of features for the two datasets.

4.4.3 Comparison with Automatic Evaluation Metrics

Figure 4.7 provides a summary of the correlation results of the proposed approach compared to state-of-the-art evaluation metrics using WMT 2007 and WMT 2013 datasets. The results show that the proposed approach has the best performance under the two datasets. The proposed approach provides an enhancement of at least 13.92% in accuracy over the best state-of-the-art techniques for the WMT 2007 datasets and at least 3.85% for the WMT 2013

Features classes	French-to-English dataset	Spanish-to-English dataset
AF	0.88	0.65
AF, CF	0.87	0.66
AF, SF	0.86	0.65
AF, CF, SF	0.92	0.83

Table 4.6: The effect of different combinations of features classes on correlation with human judgments using WMT 2007 datasets.

Features classes	French-to-English dataset	Spanish-to-English dataset
AF	0.89	0.81
AF, CF	0.90	0.72
AF, SF	0.87	0.75
AF, CF, SF	0.93	0.89

Table 4.7: The effect of different combinations of features classes on correlation with human judgments using Europarl 2005 datasets.

datasets. All techniques perform worse in WMT 2013 dataset due to the high complexity of training and testing sentences in this dataset [45].

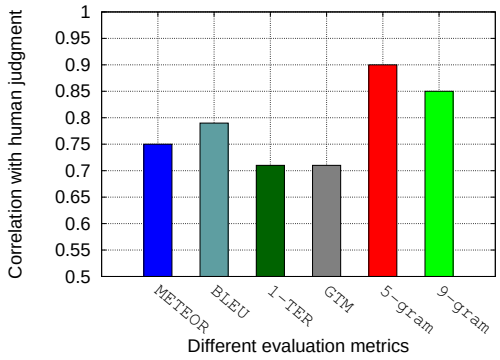
4.4.4 Effect of Weighing Features with Human Assessments

Figure 4.8 shows the effect of weighing features with normal human assessments without applying the inference model. It can be observed that the proposed approach achieves a slight improvement in accuracy for WMT 2007 French-to-English dataset The same observation applied for the Spanish-to-English dataset. The reason for these results is the inconsistency in human assessment that might happen due to the subjective nature of manual evaluation.

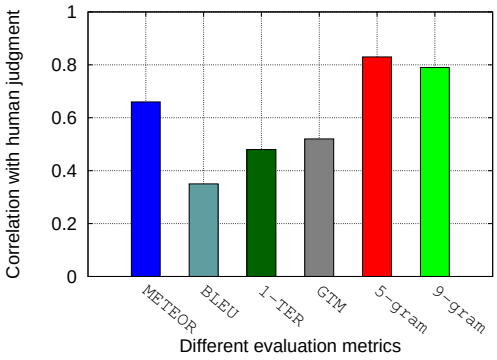
By applying the proposed inference model in section 3.8.1, the accuracy is significantly improved in the two datasets except WMT 2013 French-to-English dataset as shown in figure 4.9. The correlation values were improved by at least 6% over the values provided after weighting with normal human assessments. This reveals the importance of the inference model to solve the inconsistency issues in human assessments if these assessments exist.

4.4.5 Comparison with Quality Estimation Approaches

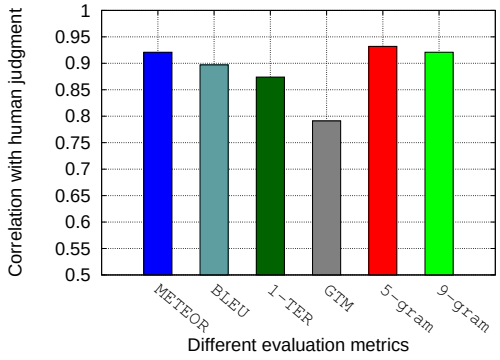
Results in section 4.4.4 show clearly that the accuracy of the proposed approach could be increased by incorporating confidence scores from credible human assessments, if these assessments are available. This finding needs to be validated by a comparison to other related



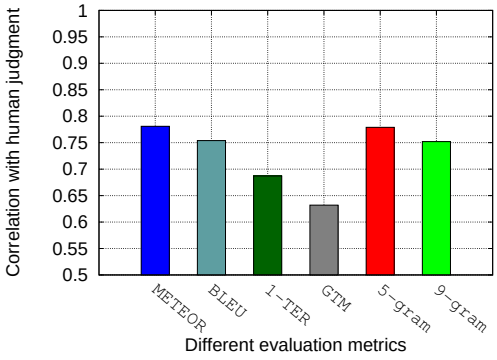
(a) French-to-English dataset from WMT 2007



(b) Spanish-to-English dataset from WMT 2007



(c) French-to-English dataset from WMT 2013



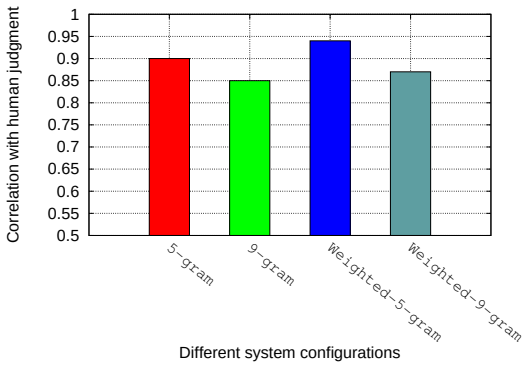
(d) Spanish-to-English dataset from WMT 2013

Figure 4.7: Comparison with traditional evaluation metrics using WMT 2007 and WMT 2013 datasets.

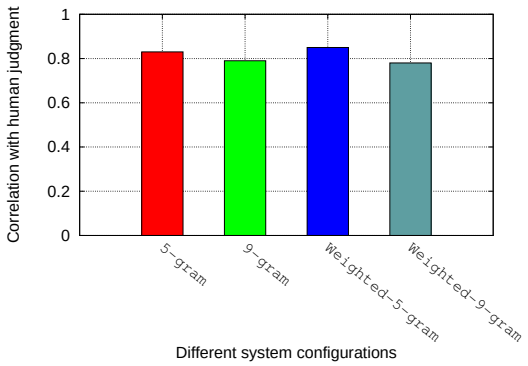
systems that depend heavily on human assessments such as quality estimation approaches.

Due to the raised interest in quality estimation, a shared task is held in WMT to discuss the recent proposed approaches in this field. This task became in action since 2012. We used WMT 2013 shared task for comparison because the official results of this task were reported in [49]. In WMT 2013, the shared task consisted of four subtasks. Each subtask provided specific datasets, and evaluated submissions from different systems using spearman correlation with human judgments. Unfortunately, these subtasks focused on Spanish-to-English datasets only. The winners of the shared task in WMT 2013 are FBK-UEDIN [44], CNGL [46] and CMU [47] systems.

Figure 4.10 shows the correlation with human judgments of the proposed approach compared to the winners of the shard task in WMT 2013. It can be seen that the proposed approach provides an enhancement of at least 12% in accuracy over the best winner.

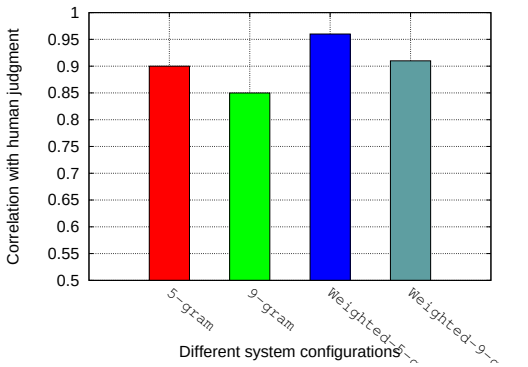


(a) French-to-English dataset from WMT 2007

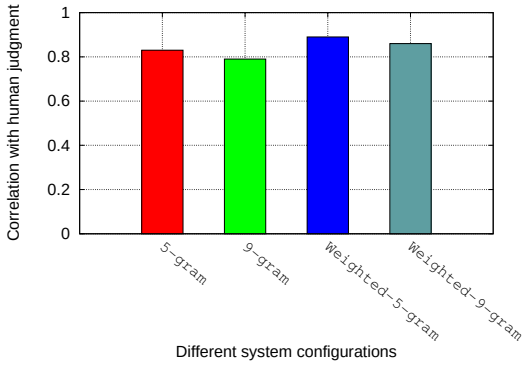


(b) Spanish-to-English dataset from WMT 2007

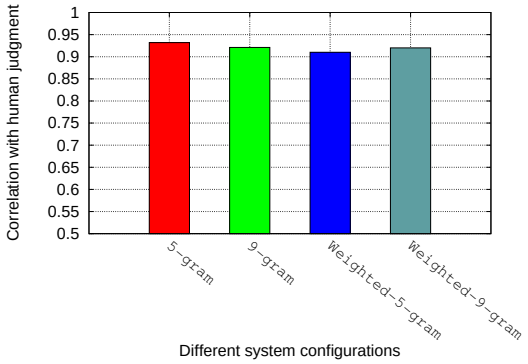
Figure 4.8: Effect of weighing features with normal human scores using WMT 2007 and WMT 2013 datasets.



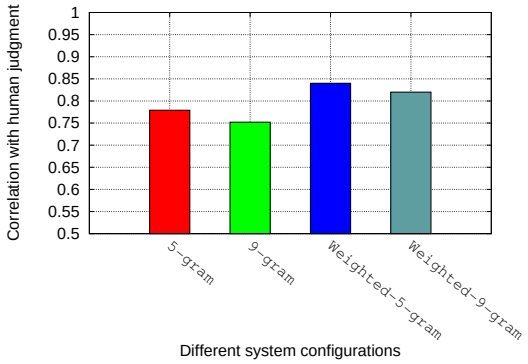
(a) French-to-English dataset from WMT 2007



(b) Spanish-to-English dataset from WMT 2007



(c) French-to-English dataset from WMT 2013



(d) Spanish-to-English dataset from WMT 2013

Figure 4.9: Effect of weighing features with human confidence scores from the proposed inference model using WMT 2007 and WMT 2013 datasets.

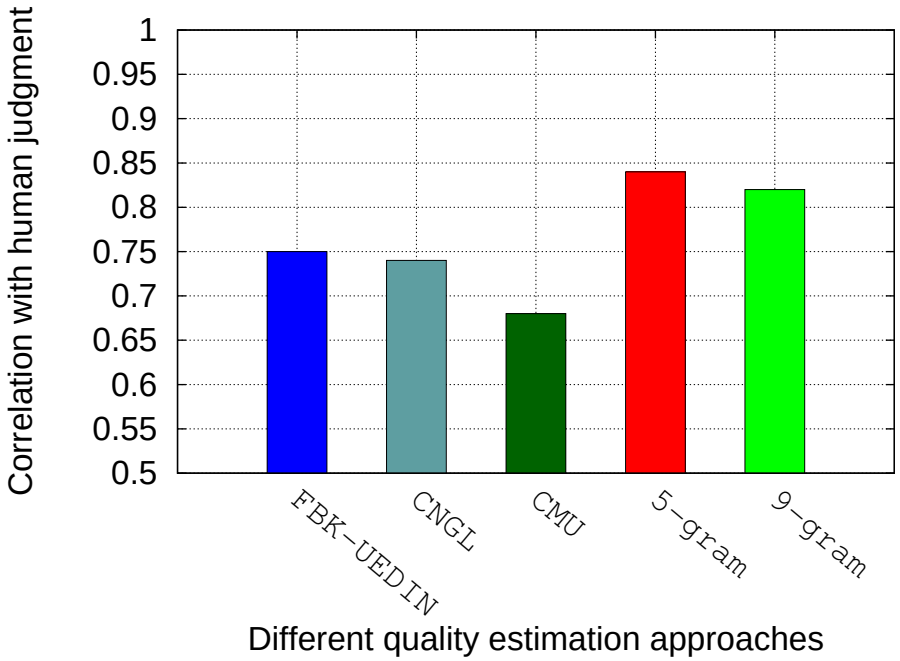


Figure 4.10: Comparison with quality estimation approaches using Spanish-to-English dataset from WMT 2013.

4.5 Conclusion

In this chapter, we presented the experimental evaluation of the proposed approach. First, we listed the evaluation metrics used throughout the chapter. Then, the datasets used for evaluation in addition to the evaluation procedure were discussed. Finally, the detailed results of the evaluation were provided.

A brief conclusion of the proposed work and some future extensions are presented in the next chapter.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In this chapter, the conclusions of the proposed work and the experimental evaluation are summarized in section 5.1. In addition, possible directions for extending this work are discussed in section 5.2.

5.1 Conclusions

In this thesis, the major shortcomings of current MT evaluation approaches were addressed. These shortcomings were believed to reduce the accuracy of evaluation. These shortcomings were mainly due to using superficial automatic evaluation metrics which may result in low accuracy. In addition, quality estimation approaches were found to depend mainly on human assessments which are scarce and have inconsistency issues.

In order to address these drawbacks, two main contributions were proposed. First, a novel set of linguistic and data-driven features for evaluating translations was provided using alignment-based models and statistics built from parallel corpora. In addition, a probabilistic inference model was proposed to infer the credibility of given human assessments by learning uncertainties in human scores and identifying bad judgments to be discarded or re-examined. After all, an aggregation formula is designed to integrate generated scores from features with confidence scores learnt from the probabilistic inference model.

To evaluate the proposed approach, experiments were conducted over French-to-English and Spanish-to-English parallel corpora. Through these experiments, the performance of the features set in addition to the proposed inference model were analyzed and compared to automatic metrics and state-of-the-art quality estimation approaches. The results of the

experiments showed that the proposed approach outperforms its counterparts.

5.2 Future Work

Several directions can be considered for extending the work presented in this thesis:

- **Heterogeneous Language-pairs**: It is possible that the performance of linguistic features will vary across different language-pairs depending on the nature of these languages, the proposed approach has not been tested enough for this (e.g. Arabic-English). However, if the phrase-to-phrase alignment model is specified before, it should be feasible to extract suitable features for this specific model because our approach is language-independent. Thus, we can build alignment models for any language-pair during the offline phase, and run the proposed system for evaluation without significant modifications. In other words, the approach would extend to all language-pairs from which initial alignment models have been built.

- **Re-inference Footprint**: The proposed approach relies on a minimal number of variables that require re-inferring periodically. This is expected to limit the offline phase overhead. Moreover, the approach, can incrementally re-infer variables, perhaps depending on the density of available users in the community. For example, if a community of users has numerous active voters, the system could avoid inferring variables related to the density of users. The current work has not addressed such optimization these are part of our ongoing work.

- **Scalability Testing**: A complete system needs to be tested over a larger scale, in terms of the number of users and dataset size. Although we provided large-scale evaluation for the proposed set of features using Europarl datasets, we couldn't do the same for the inference model. The reason for that is the difficulty to obtain a large scale dataset annotated with human assessments. This can be partially solved by providing motivating applications for judges to evaluate large batches of translations.

BIBLIOGRAPHY

[1] C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan, "Findings of the 2011 Workshop on Statistical Machine Translation," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 22–64. [Online]. Available: http://www.aclweb.org/anthology/W11-2103

[2] P. Koehn, *Statistical Machine Translation*, 1st ed. New York, NY, USA: Cambridge University Press, 2010.

[3] S. Pado, M. Galley, D. Jurafsky, and C. D. Manning, "Robust Machine Translation Evaluation with Entailment Features," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, August 2009, pp. 297–305. [Online]. Available: http://www.aclweb.org/anthology/P/P09/P09-1034

[4] P. Dirix, I. Schuurman, and V. Vandeghinste, "METIS-II: Example-based Machine Translation using Monolingual Corpora - System Description," in *In Proceedings of The 2nd Workshop on Example-based Machine Translation*, 2005.

[5] J. Denero, S. Kumar, C. Chelba, and F. Och, "Model Combination for Machine Translation," in *In Proceedings NAACL-HLT*, 2010.

[6] J. Slocum, "A Survey of Machine Translation: Its History, Current Status, and Future Prospects," *Comput. Linguist.*, vol. 11, no. 1, pp. 1–17, Jan. 1985. [Online]. Available: http://dl.acm.org/citation.cfm?id=5615.5616

[7] F. J. Och, "Statistical Machine Translation: From Single-Word Models to Alignment Templates," *RWTH Aachen University, Technical Report*, 1985. [Online]. Available: http://www-

i6.informatik.rwth-aachen.de/PostScript/InterneArbeiten/Och_SMT_From_Single-Word_Models_to_Alignment_Templates_Dissertation_08Oct2002.pdf

[8] B. Thouin, "The METEO System," in *Proceedings of Practical Experience Of Machine Translation*. In Lawson, V., editor, 1982.

[9] A. Drozdek, "Interlingua in Machine Translation," in *Proceedings of the 17th Conference on ACM Annual Computer Science Conference*, ser. CSC '89. New York, NY, USA: ACM, 1989, pp. 434–434. [Online]. Available: http://doi.acm.org/10.1145/75427.1030260

[10] U. Hiroshi and Z. Meiying, "Interlingua for Multilingual Machine Translation," in *Proceedings of the MT Summit IV*, 1993.

[11] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, Jun. 1993. [Online]. Available: http://dl.acm.org/citation.cfm?id=972470.972474

[12] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180. [Online]. Available: http://dl.acm.org/citation.cfm?id=1557769.1557821

[13] C. S. Fordyce, "Overview of the IWSLT 2007 Evaluation Campaign," in *Proceedings of International Workshop on Spoken Language Translation, Trento*, 2007.

[14] Y. Liu, Q. Liu, and S. Lin, "Tree-to-string Alignment Template for Statistical Machine Translation," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ser. ACL-44. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 609–616. [Online]. Available: http://dx.doi.org/10.3115/1220175.1220252

[15] W. S. Bennett and J. Slocum, "The LRC Machine Translation System," *Comput. Linguist.*, vol. 11, no. 2-3, pp. 111–121, Apr. 1985. [Online]. Available: http://dl.acm.org/citation.cfm?id=1187874.1187877

[16] K.-H. Chen and H.-H. Chen, "A Hybrid Approach to Machine Translation System Design," in *Computational Linguistics and Chinese Language Processing*, 1996.

[17] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, J. D. Lafferty, and R. L. Mercer, "Analysis, statistical transfer, and synthesis in machine translation," in *In Proceedings of The Fourth International Conference on Therotical and Methodological Issues In Machine Translation*, 1992, pp. 83–100.

[18] J. Yamron, J. Baker, P. Bamberg, H. Chevalier, T. Dietzel, J. Elder, F. Kampmann, M. Mandel, L. Manganaro, T. Margolis, and E. Steele, "LINGSTAT: An Interactive, Machine-aided Translation System," in *Proceedings of the Workshop on Human Language Technology*, ser. HLT '93. Stroudsburg, PA, USA: Association for Computational Linguistics, 1993, pp. 191–195. [Online]. Available: http://dx.doi.org/10.3115/1075671.1075714

[19] A. B. Phillips, "Cunei: Open-Source Machine Translation with Relevance-Based Models of Each Translation Instance," *Machine Translation*, vol. 25, no. 2, pp. 161–177, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1007/s10590-011-9109-6.

[20] E. Hovy, M. King, and A. Popescu-belis, "Principles of Context-based Machine Translation Evaluation," *Machine Translation*, vol. 16, pp. 1–33, 2002.

[21] R. Cole, Ed., *Survey of the State of the Art in Human Language Technology*. New York, NY, USA: Cambridge University Press, 1997.

[22] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan, 2005, pp. 65–72.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318. [Online]. Available: http://dx.doi.org/10.3115/1073083.1073135.

[24] C. Callison-Burch, M. Osborne, and P. Koehn, “Re-evaluating the Role of BLEU in Machine Translation Research,” in *In EACL*, 2006, pp. 249–256.

[25] P. Koehn and C. Monz, “Manual and Automatic Evaluation of Machine Translation Between European Languages,” in *Proceedings of the Workshop on Statistical Machine Translation*, ser. StatMT ’06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 102–121. [Online]. Available: http://dl.acm.org/citation.cfm?id=1654650.1654666

[26] J. Turian, L. Shen, and I. D. Melamed, “Evaluation of Machine Translation and Its Evaluation,” in *In Proceedings of MT Summit IX*, 2003, pp. 386–393.

[27] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A Study of Translation Edit Rate with Targeted Human Annotation,” in *In Proceedings of Association for Machine Translation in the Americas*, 2006, pp. 223–231.

[28] G. Doddington, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics,” in *Proceedings of the second international conference on Human Language Technology Research*, ser. HLT ’02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 138–145. [Online]. Available: http://dl.acm.org/citation.cfm?id=1289189.1289273.

[29] C.-Y. Lin and F. J. Och, “Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics,” in *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ser. ACL ’04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. [Online]. Available: http://dx.doi.org/10.3115/1218955.1219032.

[30] ——, “ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation,” in *Proceedings of the 20th international conference on Computational Linguistics*, ser. COLING ’04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. [Online]. Available: http://dx.doi.org/10.3115/1220355.1220427.

[31] S. Corston-Oliver, M. Gamon, and C. Brockett, “A Machine Learning Approach to The Automatic Evaluation of Machine Translation,” in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ser. ACL ’01. Stroudsburg, PA, USA: Association for Computational Linguistics, 2001, pp. 148–155. [Online]. Available: http://dx.doi.org/10.3115/1073012.1073032.

[32] J. Gimnez and E. Amig, “IQMT: A Framework for Automatic Machine Translation Evaluation,” in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, May 2006, pp. 685–690.

[33] D. A. Jones and G. M. Rusk, “Toward a Scoring Function for Quality-Driven Machine Translation,” in *Proceedings of the 18th conference on Computational linguistics - Volume 1*, ser. COLING ’00. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 376–382. [Online]. Available: http://dx.doi.org/10.3115/990820.990875.

[34] J. Hutchins, “Retrospect and Prospect in Computer-Based Translation,” in *Proceedings of MT Summit VII, Singapore*, 1999.

[35] L. Shen, J. Xu, B. Zhang, S. Matsoukas, and R. Weischedel, “Effective Use of Linguistic and Contextual Information for Statistical Machine Translation,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, ser. EMNLP ’09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 72–80. [Online]. Available: http://dl.acm.org/citation.cfm?id=1699510.1699520.

[36] M.-Y. Yang, S.-Q. Sun, J.-G. Zhu, S. Li, T.-J. Zhao, and X.-N. Zhu, “Improvement of Machine Translation Evaluation by Simple Linguistically Motivated Features,” *J. Comput. Sci. Technol.*, vol. 26, no. 1, pp. 57–67, Jan. 2011. [Online]. Available: http://dx.doi.org/10.1007/s11390-011-1111-1.

[37] D. Chiang, Y. Marton, and P. Resnik, “Online Large-margin Training of Syntactic and Structural Translation Features,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP ’08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 224–233. [Online]. Available: http://dl.acm.org/citation.cfm?id=1613715.1613747.

[38] A. L.-F. Han, Y. Lu, D. F. Wong, L. S. Chao, L. He, and J. Xing, “Quality Estimation for Machine Translation Using the Joint Method of Evaluation Criteria and Statistical Modeling,” in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 365–372. [Online]. Available: http://www.aclweb.org/anthology/W13-2245

[39] R. Soricut and A. Echihabi, "Trustrank: Inducing Trust in Automatic Translations Via Ranking." in *ACL*, J. Hajic, S. Carberry, and S. Clark, Eds. The Association for Computer Linguistics, 2010, pp. 612–621.

[40] K. Owczarzak, Y. Graham, and J. van Genabith, "Using F-structures in Machine Translation Evaluation," in *Proceedings of the LFG07 Conference*, Stanford, CA, 2007, pp. 383–396.

[41] L. Specia, "Exploiting Objective Annotations for Measuring Translation Post-editing Effort," in *15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium, May 2011, pp. 73–80.

[42] Y. He, Y. Ma, J. van Genabith, and A. Way, "Bridging SMT and TM with Translation Recommendation," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 622–630. [Online]. Available: http://www.aclweb.org/anthology/P10-1064.

[43] L. Specia, N. Hajlaoui, C. Hallett, and W. Aziz, "Predicting Machine Translation Adequacy," in *Proceedings of the 13th Machine Translation Summit*, Xiamen, China, September 2011, pp. 513–520.

[44] J. G. Camargo de Souza, C. Buck, M. Turchi, and M. Negri, "FBK-UEdin Participation to the WMT13 Quality Estimation Shared Task," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 352–358. [Online]. Available: http://www.aclweb.org/anthology/W13-2243

[45] M. Macháček and O. Bojar, "Results of the WMT13 Metrics Shared Task," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 45–51. [Online]. Available: http://www.aclweb.org/anthology/W13-2202

[46] E. Biici, D. Groves, and J. van Genabith, "Predicting Sentence Translation Quality Using Extrinsic and Language Independent Features," *Machine Translation*, vol. 27, no. 3-4, pp. 171–192, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10590-013-9138-4

[47] S. Hildebrand and S. Vogel, "MT Quality Estimation: The CMU system for WMT'13," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 373–379. [Online]. Available: http://www.aclweb.org/anthology/W13-2246

[48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[49] O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2013 Workshop on Statistical Machine Translation," in *Eighth Workshop on Statistical Machine Translation*, ser. WMT-2013, Sofia, Bulgaria, 2013, pp. 1–44. [Online]. Available: http://www.aclweb.org/anthology/W13-2201

[50] M. Gamon, A. Aue, and M. Smets, "Sentence-level MT Evaluation Without Reference Translations: Beyond Language Modeling," in *European Association for Machine Translation (EAMT)*, 2005.

[51] M. Paul, A. Finch, and E. Sumita, "Reducing Human Assessment of Machine Translation Quality to Binary Classifiers," in *In Proceedings of 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, 2007, pp. 154–162.

[52] D. Coughlin, "Correlating Automated and Human Assessments of Machine Translation Quality," in *Proceedings of MT Summit IX*, 2003, pp. 63–70.

[53] J. S. Albrecht and R. Hwa, "Regression for Machine Translation Evaluation at the Sentence Level," vol. 22, no. 1-2. Hingham, USA: Kluwer Academic Publishers, Mar. 2008, pp. 1–27. [Online]. Available: http://dx.doi.org/10.1007/s10590-008-9046-1.

[54] A. Kulesza and S. M. Shieber, "A Learning Approach to Improving Sentence-Level MT Evaluation," in *Proceedings of the 10th International Conference on Theortical and Methodological Issues in Machine Translation*, 2004.

[55] K. Duh, "Ranking vs. Regression in Machine Translation Evaluation," in *Proceedings of the Third Workshop on Statistical Machine Translation*, ser. StatMT '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 191–194. [Online]. Available: http://dl.acm.org/citation.cfm?id=1626394.1626425.

[56] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, Mar. 2003. [Online]. Available: http://dx.doi.org/10.1162/089120103321337421.

[57] M. Junczys-Dowmunt and A. Sza?, "SyMGiza++: Symmetrized Word Alignment Models for Statistical Machine Translation," in *Security and Intelligent Information Systems*, ser. Lecture Notes in Computer Science, P. Bouvry, M. K?opotek, F. Leprvost, M. Marciniak, A. Mykowiecka, and H. Rybi?ski, Eds., vol. 7053. Springer Berlin Heidelberg, 2012, pp. 379–390.

[58] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available: http://web.mit.edu/6.435/www/Dempster77.pdf

[59] T. P. Minka, "Expectation Propagation for Approximate Bayesian Inference," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, ser. UAI '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369. [Online]. Available: http://dl.acm.org/citation.cfm?id=647235.720257

[60] J. Winn and C. M. Bishop, "Variational Message Passing," *J. Mach. Learn. Res.*, vol. 6, pp. 661–694, Dec. 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1046920.1088695

[61] F. R. Kschischang, B. J. Frey, and H. A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theor.*, vol. 47, no. 2, pp. 498–519, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1109/18.910572

[62] P. Alevizos, "Factor Graphs: Theory and Applications," Ph.D. dissertation, TECHNICAL UNIVERSITY OF CRETE, 2012.

[63] T. Minka, J. Winn, J. Guiver, and D. Knowles, "Infer.net 2.5," 2012, microsoft Research Cambridge. http://research.microsoft.com/infernet.

[64] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(Meta-) Evaluation of Machine Translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*, ser. StatMT '07. Stroudsburg, PA, USA:

Association for Computational Linguistics, 2007, pp. 136–158. [Online]. Available: http://dl.acm.org/citation.cfm?id=1626355.1626373.

[65] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Conference Proceedings: the tenth Machine Translation Summit*, AAMT. Phuket, Thailand: AAMT, 2005, pp. 79–86. [Online]. Available: http://mt-archive.info/MTS-2005-Koehn.pdf.

[66] D. H. Stern, R. Herbrich, and T. Graepel, "Matchbox: Large Scale Online Bayesian Recommendations," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 111–120. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526725

[67] W. Chu and Z. Ghahramani, "Gaussian Processes for Ordinal Regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1019–1041, Dec. 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1046920.1088707

أعضاء لجنة الإشراف

أ.د. نجوى مصطفى المكى

أ.د. سهيرأحمد فؤاد بسيونى

أعضاء لجنة الحكم

أ.د. أمين أحمد شكرى

أ.د. نجوى مصطفى المكى

أ.د. سهيرأحمد فؤاد بسيونى

أ.د. صالح عبد الشكور الشهابى

رئيس مجلس القسم

أ.د. حسين حسن على

ملخص الرسالة

مقدمة من: المهندس ابراهيم أحمد ابراهيم صالح سابق

التقييم الذكى للترجمة الألية باستخدام الانسان و الألة

هذه الرسالة تحتوي علي اقتراح لطريقة تهدف إلى تقييم الترجمة الالية اعتمادا على مجموعة من الصفات اللغوية المبتكرة و مزجها بالتقييم البشرى الموثوق فيه متفاديا عيوب الطرق السابقة.

الفصل الأول يحتوى على مقدمة مختصرة عن عملية تقييم الترجمة الألية و أهميتها. بالاضافة الى ملخص للدافع الى الرسالة و أهم ما تم انجازة فيها. بالاضافة الى مقدمة مختصرة عن محتويات الفصول التالية.

الفصل الثانى يقدم تفاصيل عن المشاكل و التطبيقات المختلفة لتقييم الترجمة الألية و ايضا يقدم تغطية لبعض من الأساليب و التقنيات التى قام بها باحثون أخرون فى مجال تقييم الترجمة الألية.

الفصل الثالث يحتوى على شرح كامل للطريقة المقترحة لتقييم الترجمة الألية عن طريقة استغلال كلا من الخوارزميات الألية و الموارد البشرية. الطريقة المقترحة تعتمد بصفة أساسية على استخراج مجموعة من الصفات اللغوية بطريقة ألية من أمثلة قريبة للترجمة. بالاضافة الى ذلك يقدم نموذج استدلال فعال للاستفادة من التقيمات البشرية اذا وجدت. يحتوي هذا الفصل أيضا علي شرح لجميع مكونات النظام المقترح.

الفصل الرابع يحتوي على شرح وعرض لنتائج التجارب التي تم إجراؤها علي الطريقة المقترحة و مقارنة بين الطريقة المقترحة و الطرق التى اقترحها باحثون أخرون فى نفس المجال.

الفصل الخامس يلخص محتويات ونتائج الرسالة ويعرض عددا من الاقتراحات الخاصة بالأعمال المستقبلية التى يمكن إضافتها لموضوع الرسالة.

التقييم الذكى للترجمة الألية باستخدام الانسان والآلة

مقدمة من

المهندس / ابراهيم أحمد ابراهيم صالح سابق

للحصول على درجة

الماجستير في هندسة الحاسب و النظم

لجنة المناقشة والحكم على الرسالة	موافقون
ا.د. أميـــن أحمـــد شـــكرى	.....
ا.د. نجـــوى مصـطفى المـكى	.....
ا.د. ســـهير أحمد فؤاد بسيونى	.....
ا.د. صالح عبد الشكور الشهابى	.....

وكيل الكلية للدراسات العليا والبحوث

كلية الهندسة – جامعة الاسكندرية

أ.د. هبة وائل لهيطة

لجنة الإشراف	موافقون
أ.د. نجــوى مصــطفى المكى	.....
أ.د. سهير أحمد فؤاد بسيونى	.....

التقييم الذكى للترجمة الالية باستخدام الانسان والآلة

رسالة علمية

مقدمة للدراسات العليا بكلية الهندســة، جامعة الإسكندرية
استيفاء للدراسات المقررة للحصول على

درجة الماجستير

في
هندسة الحاسب و النظم

مقدمة من:

ابراهيم أحمد ابراهيم صالح سابق

يونيو ٢٠١٤