Universität Karlsruhe (TH) Fakultät für Informatik

Institut für Theoretische Informatik (ITI)
Lehrstuhl Prof. Dr.rer.nat. Alex Waibel

lti Carnegie Mellon

Carnegie Mellon University (CMU)
Language Technologies Institute (LTI)

interACT
research
Interactive Systems Laboratories

Building an English-to-Arabic Statistical Machine Translation's System for the ISL's Lecture Translator

Rim Helaoui

27th June 2007

Advisors:

Prof. Dr. Alex Waibel
Dr. Stephan Vogel

Contents

Abstract

Using machines to overcome language barriers have been the main concern of many researchers over the last 5 decades. Among the various efforts to build an automatic translator, the statistical approach showed to be specially promising. Based on machine learning methods, the statistical machine translation is data driven. Indeed it enables to build a translation system to any language pair with no need for language expertise.

The Interactive Systems Laboratories already works on many SMT-based projects. This work focuses on expanding the SMT Lecture Translator by building an English-To-Arabic SMT system. The system was built using training data mainly provided by IBM and was optimized thanks to numerous measures like data pre-processing and the generation of a more adequate Language model.

Chapter I

Introduction

1.1. Motivation

Our primitive ancestors solved the problem of communication through visual and acoustic signals. This, slowly, gave place to our current different natural languages and created a multilingual world. Each one subtly and proudly reflects a particular culture and some have, even, been praised for their beauty or their powerful support of thinking.

At the same time, this diversity also has been accused of dividing humanity by creating language barriers. Thus, while enhancing the communication within different communities, the natural languages intimidate the communication between the communities. It obstinately restrains the globalization and suppresses the effort of the progressing communication technologies.

The following scenarios depict common situations where imposed language barriers become a heavy burden:

A foreign tourist in China trying to decipher a „Dot not enter! “ -display panel

A foreign resident attempting, in an emergency, to describe some sudden sufferance to a doctor.

* Soldiers in foreign countries facing significant language hurdle when trying to communicate with local population and live interpreters can be hard to come by.

“Overcoming language barriers can be a matter of life or death in Iraq. Soldiers, medical personnel, and Iraqi citizens struggle to convey crucial information on a daily basis. While human translators are used in many situations, there simply aren't enough who are willing to assist in every important conversation” [1].

* Soldiers in foreign countries facing significant language hurdle when trying to communicate with local population and live interpreters can be hard to come by

* International and multilingual meetings and lecture usually resort to English as “standard” language no matter which nationalities and native languages the attending people have

* Within political unions, more specifically, the European Union, addressing this problem would be of a major importance in terms of economy and cultural diversity (European parliament)

Similar scenarios and frustrating circumstances were a strong motivation for many attempts to overcome these barriers and fulfill the dream of a common universal language. Prominent efforts in this direction are respectively Zamenhof's 1887 Esperanto, and Frege´s 1879 Begriffsschrift. These repeated efforts were, however, to no avail and the dream remained out of reach.

Another recent approach, „the machine translation“, has revealed more promising results than the design of universal languages and preserved the dream from being buried under the notion of „science fiction“. We will focus on this latter in the following sections.

1.2 Machine Translation: History and State-of-the-art

1.2.1 The beginning:

It is hard to associate a starting point to machine Translation. Ideas about mechanizing translation processes might be traced back to the seventeenth century even if realistic attempts appeared in the 20th century.
Indeed, the breakage of the German "Enigma" by the British could mark the beginning of Machine Translation's History.

Enigma was an electro-mechanical tool used during the Second World War to encode German military and diplomatic communications. A message was encoded using an agreed to (between the transmitter and receiver) set of wheels with specific starting positions. To encrypt the message, the text was simply keyed into the Enigma's keyboard and a light representing the encoded character was lit. Decoding messages was simply done by reversing the order of the wheels and running through the process again. The decoding was accomplished by knowing parts of messages and trying every possible wheel and initial position to find those parts of the message that matched. [2].

Along with these efforts, the American intelligence broke the Japanese Type A code.

This success inspired the famous Warren Weaver that translation wouldn't be more than decoding one language into another.
As expressed in his letter to cyberneticist Norbert Wiener in March 1947: "*I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text*".

1.2.2 The decade of optimism 1954-1965

Warren Weavers view of Machine Translation spread optimism during ten years in this field.
US- as well as international Machine Translation conferences took place and many researches were funded.
1954 Georgetown-IBM made a successful demonstration for Russian-English with carefully selected Russian sentences. However, the prediction that a fully capable machine translator would be operational within a couple of years quickly proved to be a delusion.

1.2.3 The ALPAC report 1966-1980

Due to the lack of progress, The ALPAC (Automatic Language Processing Advisory Committee) published its famous report concluding that Machine Translation was slower, less accurate and twice as expensive as human translation [3].

The impact of this publication was profound and early killed the academic Machine Translation research in the United States for a decade.
In Canada, France and Germany, the impact was less important and research continued resulting in some more or less useful systems. For instance:

The Systran System: English to French, installed in 1970 for the United States Air Force

The METEO System: For weather forecast translation from English to French, installed in Canada in 1977.

1.2.4 The Statistical approach and the beginning of the Rosetta project: 1980-nowadays

The idea that allowed scientist to decipher Ancient Egyptian could be compared to the new technique for machine translation also called Statistical Machine Translation. Scientist relied on the famous Rosetta Stone, which is a rock with inscriptions in three languages (a variety of Egyptian script, Greek and Egyptian glyphs). They cross-referenced all three texts and built the first model of how the Egyptian written language worked. (pic)
The statistical approach roughly works the same, where some text should be available in two languages and then submitted to statistical analysis on a computer. Correlations between the texts allow rendering statements of one language in the other one, even if they did not appear in the initial text. The most notable first projects based on this approach were
the GETA-Ariane (Grenoble)
SUSY (Saarbrücken)
DLT (Utrecht)
Rosetta (Eindhoven)
the knowledge-based project at Carnegie-Mellon University (Pittsburgh),
and two international multilingual projects: Eurotra, supported by the European Communities, and the Japanese CICC project with participants in China, Indonesia and Thailand. [4]

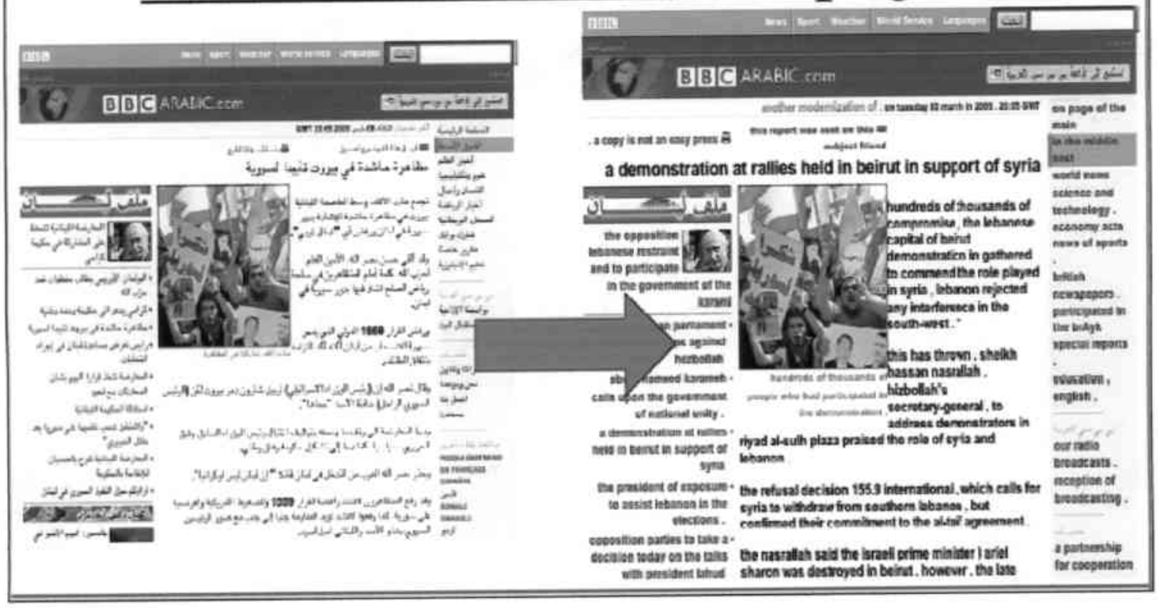
Ten years later, a group from IBM published the results of experiments on a system (Candide) based purely on statistical methods. This, along with the start of research on speech translation and involving the integration of speech recognition revealed a tremendous innovation in the field followed by a number of successful projects. On the top of which we list the collaborative JANUS project (ATR, Carnegie-Mellon University and the University of Karlsruhe), and the Verbmobil project, funded by the German government.
The large and growing cooperation resulted in a particularly impressive increase in the use of Machine Translation and translation aids in diverse applications fields

1.3 Examples of applications:

1.3.1 Previous achievements

Machine Translation software is progressively invading the current market. Its applications widely spread through different fields with the support of revolutionary technologies. Below we can see some illustrated examples issued from collaboration between the Carnegie Mellon University and the University of Karlsruhe, Germany. The projects deal respectively with web page translation based on text-to-text translation and with foreign signs translation

Text Translation (Webpages)



BBC ARABIC.com

another modernization of . on tuesday 03 march in 2005 . 20:05 GMT

. a copy is not an easy press

this report was sent on this subject friend

a demonstration at rallies held in beirut in support of syria

hundreds of thousands of compromise , the lebanese capital of beirut demonstration in gathered to commend the role played in syria , lebanon rejected any interference in the south-west . "

this has thrown , sheikh hassan nasrallah , hizbollah's secretary-general , to address demonstrators in riyad al-sulh plaza praised the role of syria and lebanon

the refusal decision 155.9 international , which calls for syria to withdraw from southern lebanon , but confirmed their commitment to the al-taif agreement

the nasrallah said the israeli prime minister l ariel sharon was destroyed in beirut . however , the late

the opposition lebanese restraint and to participate in the government of the karami

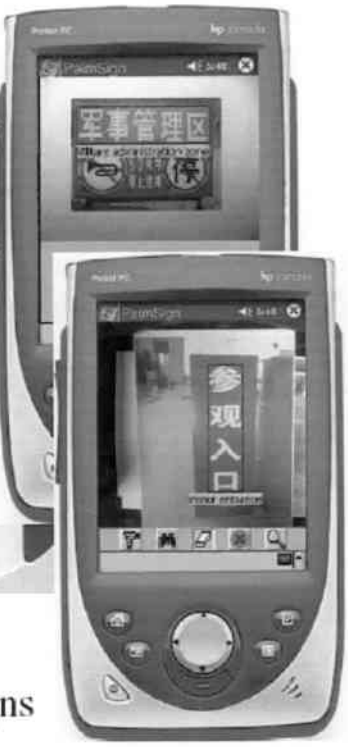
calls upon the government of national unity .

a demonstration at rallies held in beirut in support of syria

the president of exposure to assist lebanon in the elections .

opposition parties to take a decision today on the talks with president lahud

Sign Translation



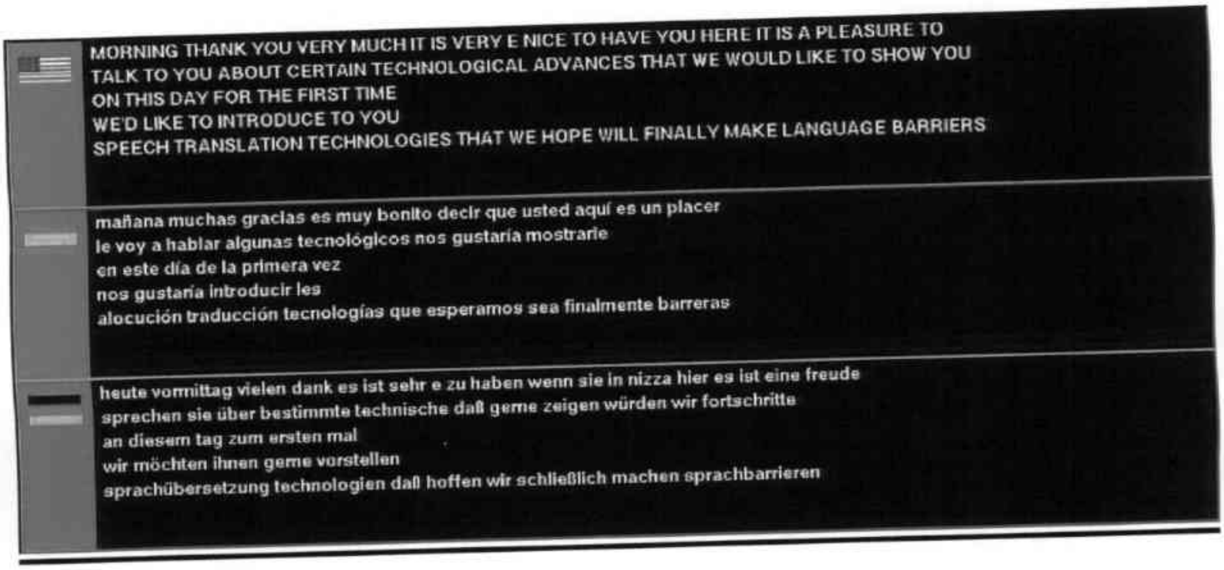
- Portable Translation of Foreign Signs

1.3.2 The Lecture Translator

This project relies on both Speech Recognition and Machine Translation to provide a real-time lecture Translation. It belongs to the Interactive systems laboratories (ISL) of the InterACT center located in Pittsburgh, Pennsylvania, US and Karlsruhe, Germany which was established in 1991 by Professor Alex Waibel.

The scientific focus of the Interactive Systems Laboratories (ISL) is on technologies that facilitate the human experience, human mutual understanding and communication. Examples of such technologies are translation, speech, language, vision technologies, multimodal and cross-modal perceptual interfaces, smart rooms or pervasive computing.
The Laboratories carry out ground breaking research on multimodality, pervasive computing, speech and improving human-computer and human-human communication. [5]

The Lecture Translator offers a bright solution to multilingual communication. It was demonstrated at Interspeech 2006, Pittsburgh, PA, USA and received the Interspeech 2006 Demo Award for "Best Presentation".
Indeed, the Lecture is supposed to be given in English and would be translated into Spanish and German simultaneously. However, for an acceptable translation quality, given talks are, so far, expected to have a limited domain.



MORNING THANK YOU VERY MUCH IT IS VERY E NICE TO HAVE YOU HERE IT IS A PLEASURE TO
TALK TO YOU ABOUT CERTAIN TECHNOLOGICAL ADVANCES THAT WE WOULD LIKE TO SHOW YOU
ON THIS DAY FOR THE FIRST TIME
WE'D LIKE TO INTRODUCE TO YOU
SPEECH TRANSLATION TECHNOLOGIES THAT WE HOPE WILL FINALLY MAKE LANGUAGE BARRIERS

mañana muchas gracias es muy bonito decir que usted aquí es un placer
le voy a hablar algunas tecnológicos nos gustaría mostrarle
en este día de la primera vez
nos gustaría introducir les
alocución traducción tecnologías que esperamos sea finalmente barreras

heute vormittag vielen dank es ist sehr e zu haben wenn sie in nizza hier es ist eine freude
sprechen sie über bestimmte technische daß gerne zeigen würden wir fortschritte
an diesem tag zum ersten mal
wir möchten ihnen gerne vorstellen
sprachübersetzung technologien daß hoffen wir schließlich machen sprachbarrieren

Chapter II

Statistical Machine Translation: an overview

2.1 What makes Machine Translation hard?

Translating from one language to another poses many challenges to the computers. Many ambiguities are supposed to be solved and significant lexical, syntactical, semantic as well as pragmatic differences are faced before the text appears in the target language.
These differences principally involve word order, word sense, pronouns, tense, and Idioms.
Let's consider the language pair English/Arabic for some illustrations, which would still apply for many other language pairs.

(Please notice that Arabic script is read from right to left.)

-Word order:

The English word order is subject-verb-object:
However Arabic allows verb-subject-object order as well as subject-verb-object

My father bought **a car**	bought my father **a car**
*ابي اشترى سيارة	اشترى ابي **سيارة**

-Word sense:
Most of the words have different meanings. Discerning the right one is necessary for a correct translation. This is strongly dependent of the context. The following examples illustrate this ambiguity:

خيار : **could mean** alternative **or** cucumber
Book: **could mean** حجز **or** كتاب

- Different tenses:

English has different tenses for the past. These tenses have different nuances that give more precise information about the time of the action, for example whether it is finite or not. When translating from Arabic to English we would renounce that since Arabic verbs have only one past tense.

مات الرجل = The man has died / The man died ?

-Idioms:

Each language reflects some culture and has its own Idioms. These Idioms, when translated literally to any other target language, would produce a senseless sentence. This hardens the machine translation task, requiring a deeper knowledge from the computer. As example we refer to the common Arabic Idiom:

زاد الطين بلة which means : made the situation worse. Yet, when translated literally outputs "increased the wetness of the clay"

The mentioned obstacles did not prevent serious researches in the Machine Translation field. Different approaches- widely varying in performance- have been adopted starting from word-to-word translation to the modern Statistical Machine Translation.

2.2 Machine Translation approaches

Machine translation has been approached in various ways over the last 50 years. Based on the adopted strategy, these methods could be assigned to six main classes. The direct approach, the transfer approach, Interlingua, statistical machine translation, controlled language and example-based translation [6].The last two classes go beyond the scope of this review.

2.2.1 The direct approach:

Also known as word-for-word translation, this approach appeared first and is primarily based on a bilingual dictionary to translate each word of the source text. The text in the target language is generated conform to the following steps:

*Preprocessing: Extract base forms from the source text with removing morphological inflections.

*Look up the translation of the obtained lemmas using a bilingual dictionary

*Generate the output text using the translated words from step two and some order-related rules of the target language

The resulting text is, however, of a low quality and usually lacks coherence since the both context and the connections between the words are ignored. This invokes the necessity of an intermediate stage before mapping between language pair.

2.2.2 The transfer approach

This approach aims to transfer the syntax to the target sentence as well as to preserve the semantic. Again the procedure is carried out in three steps:

*Analysis: the source text sentences are parsed into VPs, NPs, V, SUBJ, OBJ etc and outputted in a parse tree.

*Transfer: the parsed components of the source text are rearranged generating a new tree to match the required order of the target language.

*Generation: the nodes of the new tree are translated into the target language.

2.2.3 The Interlingua approach

Introducing an intermediate, independent and universal language between the source and the target language is the basis of this approach. The idea is to represent the semantic of the source language sentences using an abstract logical form, from which any other natural language could be easily generated.

2.2.4 Statistical approach and its advantages:

All of the previously mentioned approaches require crucial linguistic expertise and only face the machine translation challenges partially. Indeed, they still don t resolve most of the critical problems like, for example, Idioms and word ambiguities.

On this level, the Statistical approach is much more promising. Instead of focusing on giving a precise way of transforming one text from a source language to a target one, Statistical Machine Translation is rather concerned of the result itself, not the process.

It is clear that some phrases in some languages just don t have any correspondent in other ones, which is justified by the huge variety of languages as well as the cultures hiding behind them and implicitly reflected through some expressions, metaphors and concepts. Therefore we could conclude that translation is sometimes impossible. A compromise between faithfulness to the source expression as well as fluency of the target one imposes itself.

The statistical approach aims to output sentences which produce the best "product" between "fluency" and "faithfulness". Given a parallel corpora consisting of the same text in the source and the target languages as input, this learning-based approach tries to find the best parameters for some translation models which would coincide with our "faithfulness" concept.

This leads us to the following simplified expression of the outputted text:

$$\textbf{Translation} = \underset{T}{\textbf{argmax}} \ \ \textbf{Faithfulness (T, S)} * \textbf{Fluency (T).}$$

Where T stands for target language and S for source language.[7]

Thus, it should be obvious that the statistical approach offers important advantages compared to the preceding ones. It provides a solution to lexical ambiguity as well as Idiom that appear in the used training data without requiring the least linguistic knowledge neither for the source language nor for the target. It is totally independent from the language itself. It can quickly, at low cost prototype a new system to any language pair provided there is enough training data.
The next section allows for a closer look into the system and its components.

2.3 Statistical Machine Translation

2.3.1 Noisy Channel

Let us recall the famous declaration of Norbert Wiener in March 1947: “*I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text*”.

This is a key idea to the noisy channel concept and its application in the statistical machine translation domain. Norbert Wiener considered a Russian sentence as originally formulated in English than turned out to Russian after being “encoded”, or in other words, corrupted by “noise”.

For a concreter representation, consider the case of two people trying to communicate with each other in a very noisy place. Utterances could arrive in a non understandable way to the second person who would still try to obtain the delivered information. Thus this person needs to figure out how noise could alter spoken words and transform them into the received sounds and, at the same time, needs to project what kind of information, idea and sentence his partner could have thought of. These two steps could reconstruct the originally, most likely formulated sentence.

The statistical machine translation problem follows the same scheme, where translating from a sentence f, in some foreign language, to a sentence e in English, corresponds to recovering the most likely sentence e which could have been turned into f.

This approach of translating f back into e might sound odd and unnecessarily complicating the translation task. Nevertheless, this theory is needed to achieve an automatic, data driven approach, as it will be shown later.

2.3.2 Basic probability recall and Bayes Rule [8]

Adopting the preceding reasoning, also referred to as Bayesian reasoning, requires probabilistic methods to retrieve the most likely sentence from all available hypotheses. Therefore we will introduce some fundamental definitions.

We will consider the case where some text “e” in English is to be translated into the text “f” in some foreign language:

P(e) (a priori probability): the probability that the sentence “e” appeared at a certain time.

P(f / e) (conditional probability) : the probability of “f” given “e”, in other words, given that “e” has appeared, what is the probability that “f” is produced.

P (f, e) (joint probability): the probability that both of “e” and “f”.

$\underset{e}{\text{Argmax}}$ P(e/f) : the sentence “e” which would maximize the value of P(e/f).This would correspond to the most likely translation we are looking for. (See the previous subsection)

Now, based on these definitions, we arrive to the important Bayes Formal which justifies the need for the noisy channel theory to solve our automatic translation problem.
The formal connects the previous probabilities as follows:

$$P\ (e/f) = P\ (e) * P\ (\ f/e) \ /\ P(f)$$ [1]

Since e and P(f) are totally independent, then

$$\underset{e}{\operatorname{argmax}}\ P(e/f) = \underset{e}{\operatorname{argmax}}\ P(e) * P\ (f/e)$$

Thus, the best English translation "e" is the one that maximizes the value of both:

1- the probability that "e" appears and
2- the probability that, when it appears, it gets transited into "f"

Which is nothing else than the Noisy Channel Theory explained above.

2.3.3 SMT models

An obvious similarity between the Bayesian reasoning and the faithfulness/fluency concept mentioned in the previous subsection makes it clear that this latter is no more than a non-formal description of the Bayes Rule. Hence the statistical translation task involves two models; To the first one, also called Language Model, we attribute the "fluency" notion since it will provide us with the required P(e) probabilities. To the second one concords, indeed, with what we described as "faithfulness" and it will provide P(f/e) probabilities required for Translation Model. Since the two models are independent and automatically computable, we conclude that the Bayes reformulation offers a solution to a data drive, learning-based machine translation.

Language Model: aiming to guarantee a certain level of grammatical and syntactical quality as well as to give a priority to expressions which are more specific and more common in the target language.

One intuitive idea about how to build a Language Model is to create a huge database containing any sentence that ever appeared in that language as well as how many times it appeared. This would yield a certain probability P (e) for each sentence "e" in the database, yet a P (e) = 0 for each other sentence e.

Generating such a complete database being currently impossible, another approach is, hence, required. This latter should adjust the deficiency of the former idea. In other words, never assign a probability of zero to sentences that good be perfectly correct just because has never appeared in the database.

[1] $P(e, f) = P\ (\ e/\ f) * P\ (f) = P\ (f/e) * P\ (e)$

N-grams Language Models have, so far, achieved a good step in this direction. They are inspired from the human way of recognizing "good" sentences from "bad" ones, through considering the sentences as a set of substrings of n-words, also called n-grams. The idea that a set of "good" substrings often yields a "good" substring justifies this reasoning and proposes applying the conditional n-gram probability.
The computer estimates the probability that some unigram x appears, given some substring Y. The length of the considered substring (Y+ x) is equivalent to N.

As for example the probability of starting a sentence with the unigram "the" multiplied with the probability of seeing the unigram (man) after the bigram (, the) multiplied with probability of seeing the unigram "is" given the bigram (The, man) multiplied with the probability of the appearance of the unigram (dead) after the bigram (man, is), would assign a trigram model probability to the whole sentence "The man is dead".

P (The man is dead) = P (The/ - -) * P (man/ - The) * P (is / The man) * P (dead / man is) * P (-/ is dead)

Notice that this approach still enables zero probabilities. This still can be easily adjusted by introducing the smoothing coefficients. Indeed, the idea of smoothing consists in increasing the probability P (C/ A B) attributed to some trigram ABC through including the probability of the substrings BC and C. Yet, seeing the exact trigram ABC in the database should have more effect then seeing the substring BC which at its turn should have more weight than seeing the unigram C. The language model should provide the best weights to each factor.

Both data used to generate the probabilities of each n-gram as well as the particular set of parameters make that some language models are "better" than others. This induces some metric of the "goodness" of some language model compared to others. This metric is called perplexity.
The language model should thus be tested with some unseen test data and the higher P (Language Model / test data), the better it is. At this level we have again recourse to the Bayes Rule and conclude that the model, which maximizes P (Language Model / test data) is the one which maximizes P (test data / Language Model), assuming that all language models are a priory equally probable.

Consequently, evaluating a language model relies on the probability it assigns to some unseen test data. This probability, being a product of man small factors, we express the perplexity as follows:

$$\text{Perplexity} = 2^{-\log(\text{P(test data)}) / N}$$

Where N designs the number of the words in the test data and is needed for normalization purpose.

Translation Model:

The next step according to the formals introduced above is to worry about what we called "faithfulness". Indeed, we need to assign probabilities P (f /e) to the foreign language string f given a hypothesis English translation e. This is designed as Translation Model. Obviously, these probabilities cannot be assigned in the same way as Language Model does, i.e. we cannot apply:

$$P\ (f/e) = \text{count}\ (f, e)/\text{count}\ (e)$$

since there is a huge variety of sentences.

However, choosing a specific target word ei in e we can then calculate the probability t (fi / ei) as follows:

$$T (fi / ei) = count (fi, ei) / count (ei)$$

At this point we recall that the bilingual corpora is, unfortunately, not word aligned which creates an obstacle to calculating t (fi / ei) . Let us design with "a" the function which aligns our corpora and assigns to each fj from f = f1..fm the best corresponding word eaj from
e = e1..el. There are many alignments possible with different probabilities. An alignment a would be highly probable if it links words from e and f which are translations of each other.

Thus, finding out the right t (fi / ei) requires to find out the right alignment, yet the right alignment would be specified if the t (fi / ei) probabilities were available. The Estimation-maximization algorithm (EM) offers an exit to this maze and provides the needed parameters using an iterative optimization.

Please notice that the represented translation model refers to the original IBM model which covers, in fact, other parameters than t (fi / ei) so that to achieve better translation. This lies, however, beyond the scope of this report.

-Decoding:

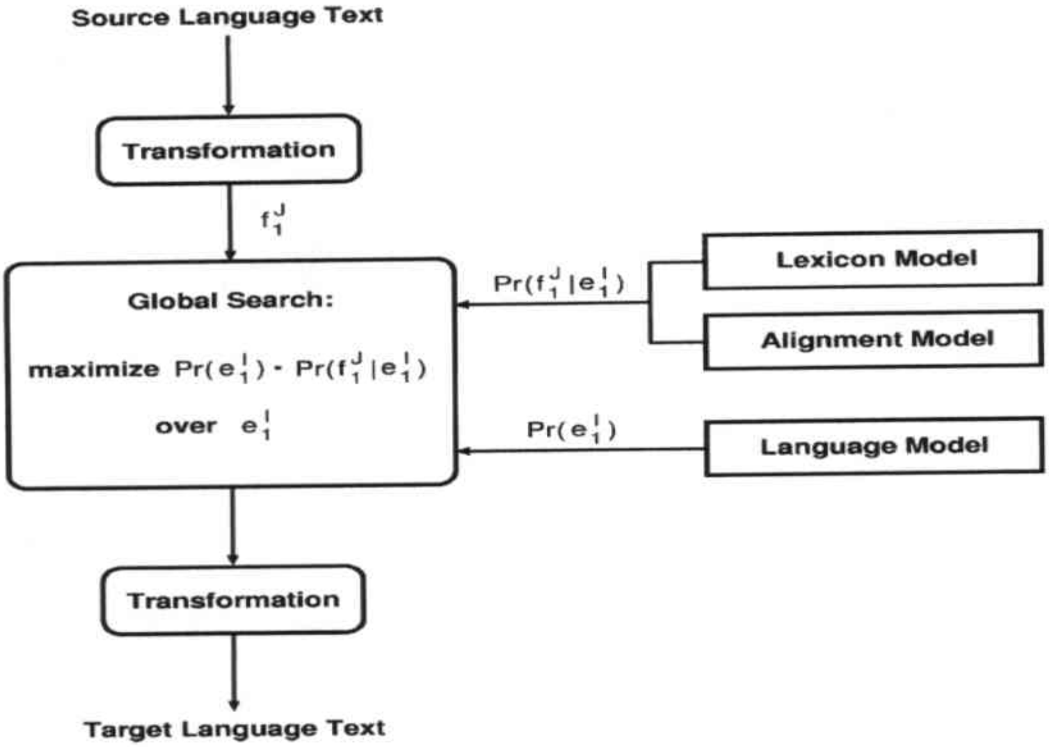
The introduced models supply us with values to compute P(e) from the monolingual corpus and P(f/e) from the bilingual one. The decoding process consists in finding the sentence "e" which maximizes the product of these terms. It implies inspecting the subset of the highly relevant sentences using particular data structures as well as heuristic search algorithms.
The next section provides some more details about the decoding process in the CMU SMT system.

-Pre-processing:

The following schema recaps the previous components and processes and represents a clear view of the translation Bayes architecture.

According to the schema, there are, obviously, two more required steps to generate the desired output. These steps are referred to as "Transformation" or pre-and-post-processing of the data, which denotes "*different input modification applied to raw training and testing texts*" [9]

Indeed, the choice of input representation is highly relevant since it affects the consistency between training and testing particularly for languages with orthographic Ambiguity. It also includes morphological preprocessing like stemming and part-of-speech (POS) and could be a key to improving the quality of the output text for some language pairs. Morphological preprocessing will be examined with more details in section four.



Schema of the translation Bayes architecture.

- Evaluation: [10]

In the purpose of automatically evaluating the output of statistical machine translation systems, some metrics have been defined. The primary measures used to rate a given system are:

BLEU: BiLingual Evaluation Understudy. [11]
It assigns a higher value to outputs which have a larger number of matching n-grams with some given references. At the same time it aims to guarantee some balance between the length of translation and the reference sentences.

A variant of the BLEU metric is the NIST. The NIST metric compares the output to some given references and calculate information gain for the matching n-grams then sum them up. It also penalizes sentences whose length significantly differs from the considered reference.

Other metrics can be used to reflect the error rate, i.e. the percentage of words to be deleted, Inserted or substituted to render the system output similar to some given reference.
We cite the mWER (Multi reference Word Error Rate), mPER (Multi reference Position independent Word Error Rate)

Chapter III

Related work

3.1 The CMU Statistical Machine Translation System: [12]

The previous section mainly focused on the original IBM translation model to give an overview to statistical Machine Translation systems. The CMU system, like many other modern systems, tries to incorporate other features and structures.

Contexts as well as local information seem to elude the IBM word-based concept. Phrase-to-phrase translation, however, outperforms these models. First using phrases to gain some lexical content and then estimating the translation model parameters output a better and more robust translation. This makes more sense if we considered the different words orders used in each language as well as some language specific expressions and idioms.
At this level, the translation quality is strongly dependent on the automatic phrases extraction, i.e. whether the extracted phrase pairs are translations of each other.

The CMU SMT system relies on word-to-word alignment in order to find phrase-to-phrase translations from a bilingual corpus. More specifically, it uses the Viterbi path generated from the HMM-based alignment model (Vogel et al., 1996) "mapping source phrases ranging from positions $j1$ to $j2$ the corresponding target phrase is given by

$$\mathbf{i_{min} = min_j \{i = a(j)\}}\ \textbf{and}\ \mathbf{i_{max} = max_j \{i = a(j)\}},\ \textbf{where}\ j = j1 \quad j2 \text{ “}$$

Once the phrases are extracted, the transducer can be built.

*"We define a transducer as a set of translation pairs generated by (*phrase extraction*) methods as well as by alternative knowledge sources such as manual dictionaries. Each translation pair has the form*

Label # Source # Target # Probability

where the label can be used to build hierarchical transducers" (Vogel et al., 2000)

"The transducers are organized as prefix trees over the source side, with translations and translation probabilities attached to the final nodes. This allows for efficient search"(Vogel et al., 2000) during the decoding.

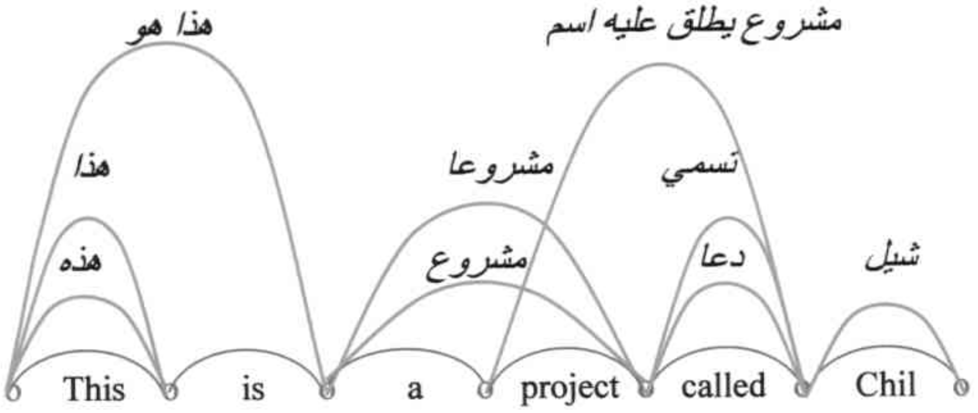
Applying all the transducers builds the translation lattice over the source words. This is done either:

Online: first by loading the bilingual corpus and indexing the source side using a suffix array. Then all n-grams of the test data are located and the aligned target phrase is added to the translation lattice

Or Offline: by loading a phrase table which is generated during the training.
For an efficient search of the matching test n-grams, the source side of the generated phrase table is indexed as prefix tree .Finally each translation associated to any of those matching sequences of words is added to the translation lattice.

Please notice that matching a sequence of words through a transducer with a part of the test sentence can start at each position in the sentence (also called a node).
Expansion of a hypothesis, i.e. moving over an edge in the lattice implies the same expansion in the transducer tree.

The following scheme: illustrates an example of translation Lattice:



Translation lattices, indeed, don't inform only about the target translation but also carry a number of model scores. The lattices´ best path along with the applied language model results in the translation hypothesis, i.e. a partial translation and a score. The whole process could be observed in the following sequence taken from a Log file generated during the decoding stage:

```
TRANSLATION_RESULT_SENTENCE 17 this is a project called chil
Build Translation Lattice
- Apply Transducer
CDecoderBLM::BuildTranslationLattice
check existing translations from phrase table
Search phrase pairs from corpus
CPhraseSearchMS::FindTranslations
BeamFactor = 2 MaxTranslations = 10
All Edges:
0 edge (0-1) Word=this Score=0 StartFrame=0 EndFrame=0  []
1 edge (1-2) Word=is Score=0 StartFrame=0 EndFrame=0  []
2 edge (2-3) Word=a Score=0 StartFrame=0 EndFrame=0  []
3 edge (3-4) Word=project Score=0 StartFrame=0 EndFrame=0  []
4 edge (4-5) Word=called Score=0 StartFrame=0 EndFrame=0  []
5 edge (5-6) Word=chil Score=0 StartFrame=0 EndFrame=0  []
```

```
6, 0.404661, 0.692704]
7 edge (0-1) Word=@PESA-MS Score=0 StartFrame=0 EndFrame=0 { this #  { 2.71128 # هذه
0.386648 ,0.0833814 ,0.0185113 ,0.00921323],
 0.982861, 1.23067]
[...]
9 edge (0-2) Word=@PESA-MS Score=0 StartFrame=0 EndFrame=0 { this is #  { 2.10477 # هذا هو
0.114648 ,0.258672 ,0.352402 ,0.075235],
 0.931057, 2.47753]
[...]
0 edge (1-4) Word=@PESA-MS Score=0 StartFrame=0 EndFrame=0 { is a project #  { 7.91151 # مشروع
0.004 ,7.48119 ,0.252947 ,1.51393]
32099, 10.492, 3.99015]
]
[...]
8 edge (2-4) Word=@PESA-MS Score=0 StartFrame=0 EndFrame=0 { a project #  { 4.63811 # مشروع
0.0008 ,1.25958 ,0.0805774 ,0.234997]
85378, 5.03965, 2.66053]
119 edge (2-4) Word=@PESA-MS Score=0 StartFrame=0 EndFrame=0 { a project #  { 6.12859 # مشروعا
0.02688 ,1.19146 ,0.0331146 ,0.12093]
45, 2.78927, 8.09553]
[...]
1 edge (3-5) Word=@PESA-MS Score=0 StartFrame=0 EndFrame=0 { project called #  مشروع يطلق عليه
0.00132 ,0.171126 ,0.0766023] { 5.50491 # اسم
596, 1.19064, 2.3772, 7.19294]
[..]
3 edge (4-5) Word=@PESA-MS Score=0 StartFrame=0 EndFrame=0 { called #  { 3.40549 # دعا
0.0326 ,0.0734717 ,0.0113534 ,0.0131759]
424, 0.921596, 2.35325]
154 edge (4-5) Word=@PESA-MS Score=0 StartFrame=0 EndFrame=0 { called #  { 4.07176 # تسمي
0.00 ,0.000296696 ,0.0104732 ,0.0326287]
[...]
1 edge (5-6) Word=@PESA-MS Score=0 StartFrame=0 EndFrame=0 { chil #  { 5.65356 # شيل
0.00191072 ,0.20568 ,0.992765 ,0.406538],1.82186, 2.22481]
[...]

Number of final hypotheses is: 112
NBest = 500
Build N-Best List
Unique translations added to n-best list = 500  size of NBestTranslations = 500
Normalize Translation Results
Display Nbest Translations
TRANSLATION_RESULT_HYP   17 1 وهذا المشروع دعا شيل
```

Sequence taken from a Log file generated during the decoding stage

A third step in the CMU system decoding process after the word-to-word translations and the phrase-to-phrase translations deals with specific information like Named Entities (NE) translation tables which include named persons, locations, and organizations.

As last step comes the optimization. Using specific parameters and models results in a specific translation score. So in order too find the best translation with the Minimum Error Training, these models have to be weighted with an optimal set of weights reflecting the reliability of a model compared to the others. Thus the best scaling factors which maximize BLEU and NIST scores are obtained. [13]

3.2. The CMU SMT environment and tools:

The SMT Group at Carnegie Mellon University, which is a part of the InterACT research center established by Professor Alex Waibel, leads ones of the most promising researches in the field of Statistical Machine Translation. The group has sixteen members working under the supervision of Dr. Stephan Vogel. Weekly meetings, an efficient use of available resources and CPUs through the telnet/ssh client Putty and the incorporation of the condor job management system has guaranteed a successful teamwork and outstanding results.

The SMT group has been participating in many international projects like StrDust, TcStar, STEEM, TransTac and the GALE Project [1].Each project uses particular training and test data. For this work, the data for GALE evaluation 2006 was used as training data. The English-Arabic parallel corpus is chiefly composed of news portions and some UN data and reaches the size of over than 1M words (more than 1GB memory). For some particular purposes and experiences, there was a need for a smaller corpus. The data for the IWSLT evaluation 2006, which contains less than 200000 words provided by ATR, was convenient.

Apart from the data, the CMU Statistical Translation Toolkit; TranslateMM is available for the Group's members. The toolkit is composed of a decoder, languages models, confusion network translation and phrase Table pruning. The Group also contributes to a set of tools usually written in C/C++, in Java, in Perl or as Unix-scripts needed for some tasks like, for instance, indexing corpus using suffix array, analyzing the corpus, Tagging Named-Entities and Pre-processing the data.

Some of the pre-processing tools will be introduced, in the next section, as a part of the related work of Sanjika Hewavitharana who is working on the Arabic-English system.

[1] *GALE (Global Autonomous Language Exploitation): The goal of the GALE program is to develop and apply computer software technologies to absorb, analyze and interpret huge volumes of speech and text in multiple languages. The LTI's machine translation efforts within this program are aimed at generating the highest-quality translation by improving statistical, example-based, and transfer-based machine translation systems and producing a multi-engine combination of their outputs which outperforms any single translation system used in the combination."* [14]

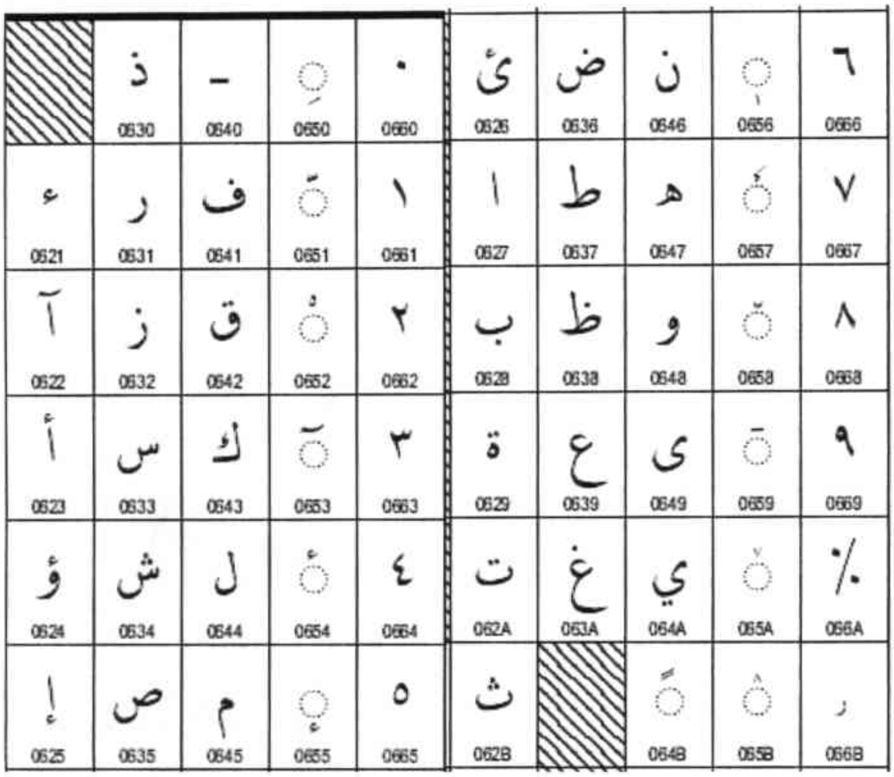
3.3 Overview of the Arabic to English system created by Sanjika Hewavitharana:

3.3.1 The Arabic language challenge

One of the main assets of statistical machine translation systems is that it could be built for any language pair. However, the confronted complexity distinctly varies from one language to another. Arabic presents a unique challenge to statistical machine translation systems. These systems, being data driven, a huge amount of data in Arabic should be available and operable by the computer. Statistical machine translation system experience Arabic data shortage especially for some specific domains, but they also should heed and deal with the following particularities of the Arabic language.

The Arabic script:

Unlike English and other Latin-based languages, the Arabic script is written, read and encoded from right to left. CMU STM systems use the following Unicode Transformation Format (utf-8) for its Arabic data:



	ذ 0630	ـ 0640	◌ِ 0650	٠ 0660	ئ 0626	ض 0636	ن 0646	◌ٖ 0656	٦ 0666
ء 0621	ر 0631	ف 0641	◌ّ 0651	١ 0661	ا 0627	ط 0637	ه 0647	◌ٗ 0657	٧ 0667
آ 0622	ز 0632	ق 0642	◌ْ 0652	٢ 0662	ب 0628	ظ 0638	و 0648	◌٘ 0658	٨ 0668
أ 0623	س 0633	ك 0643	◌ٓ 0653	٣ 0663	ة 0629	ع 0639	ى 0649	◌ٙ 0659	٩ 0669
ؤ 0624	ش 0634	ل 0644	◌ٔ 0654	٤ 0664	ت 062A	غ 063A	ي 064A	◌ٚ 065A	٪ 066A
إ 0625	ص 0635	م 0645	◌ٕ 0655	٥ 0665	ث 062B		◌ً 064B	◌ٛ 065B	٫ 066B

According to this encoding, some Arabic letters are assigned two different codes. This denotes another particularity of the Arabic language, which should be taken in consideration at some stages, for instance during post-processing.

Indeed, some Arabic letters do have different aspects depending on their position in the word, i.e. some letters would not look the same at the beginning of the word as in the middle, at the end of word or on their own:

"لعبة"   <<   "تغيير "

"a game"   "a change "

Arabic letters do have different aspects depending on their position in the word

A second facet of the Arabic orthographic ambiguity is the presence of diacritics. Those are present below or above each character in the word and are used to represent short vowels as well as doubled consonants. Those diacritics, as a result, could give different meanings to the same sequence of letters generating, thus, different words.

↗ يقتل = To kill

يقتل   ?

↘ يقتل = To be killed

Being rather optional, the Arabic diacritics are usually totally or partially absent with a few exceptions like some religious books as well as textbooks for children and foreign learners. Therefore, distinguishing the right pronunciation and the correct word purely relies on the context and is usually performed in an automatic way by native speakers. This orthographic ambiguity is even enriched with an occasional, yet less often, absence of some other letters like the symbol "ء " called "Hamza". The Hamza could appear upon or below the "Alif" symbol "ا " as follows: "أ "and "إ ". The absence of the Hamza is rather common which leaves a bare Alif symbol.

All this not only increases the complexity of some processing steps like transliteration, which is employed in some translation systems (i.e. systems using the Buckwalter[1] tool), but could also lead to severe inconsistency between training and test data as well as within the same training corpus. As a consequence, the data should be a subject to some "manipulation" or more precisely pre-processing and normalization. These required and conventional steps are presented in the next subsection.

[1] Buckwalter: **Arabic transliteration** was developed at Xerox by Tim Buckwalter in the 1990s. It is an ASCII only transliteration scheme, representing Arabic orthography strictly one-to-one, unlike the more common romanization schemes that add morphological information not expressed in Arabic script. (Wikipedia)

3.3.2 Pre-processing and normalization

For the Arabic side of CMU SMT Arabic-English system, Sanjika Hewavitharana uses the IBM normalization. This mainly consists of the following rules suggest by Yaser Al-Onaizan (IBM Research)

- Hamza normalization: maddah-n-alef, hamza-on-alef, hamza-under-alef mapped to bare alef

آ/أ/إ → ا

- Alif maqSuura mapped to yaa: ى → ي

- replace yaa followed by hamza with hamza on kursi (yaa): ىٔ → ي

- Deleting diaclitics

- Punctuation normalization since Arabic punctuation is slightly different from English punctuation.

- Digits normalization

Along with some other rules which deal, for instance, with rarely-used characters or Arabic Presentation Form-B to Logical Form.

3.3.3 The Arabic-English system using IWSLT data:

A starting point to the lecture translator was the Arabic-English system built using the IWSLT04 test set and 20000 Arabic-English BTEC sentences as training data.
Here is a brief presentation of some of the main parameters characterizing the system:

And assigns *PhraseAlignStretchFactor* = 1.2

This is justified with the richer morphology of the Arabic language compared to English. Please notice that the stretch factor is obtained by dividing the number of the words in the target side by that of the source side.

The System uses a 6-grams suffix array *language model* using the indexed English side of the parallel corpus. Building the suffix array is done using scripts written by Joy (Ying Zhang) with *CorpusName* and *CorpusSize (millions of words)* as input and the following files as output:

*.id_voc (Stores the vocabulary built for the corpus)

*.sa_corpus_v6
*.sa_offset_v6 } (Indexing the corpus)
*.sa_suffix_v6

Under *Lexicon* and *reverse Lexicon* parameters, the paths of the lexica obtained during the training are given. The Lexicon (and respectively the reverse lexicon) would have the following form: source_word target_word probability (respectively target_word source_word probability)

العذاب	misery	4.068583e-05
العذاب	torture	5.665420e-03
العذاب	suffer	2.254165e-04

A sample from the reverse lexicon

Next come the suffix array and corpus respectively under *SourceFileSA*, *SourceFile* and *TargetFile*. As explained above, these parameters are needed for online phrase extraction or for the generation of the phrase table.

In the case of an offline decoding, the path to the phrase table should be given under the *PhraseTableFileName* parameter.

At last, optimization's and evaluation's parameters should be specified, i.e.:

MEROptimize: activates (respectively deactivates) the minimum error training.
IterationLimit: specifies the number of iterations for the minimum error training.
ScoringMetric: specifies the adopted evaluation metric.
NumReferences, *ReferenceList*: respectively specify how many references are used for the evaluation and their path.

NormalizingScript: this refers to the post-processing script, which is an important step comparable to the pre-processing one introduced above. Indeed, the output data should be edited to match, as much as possible, the reference text. The normalizing script used in the IWSLT translation system was written by Sanjika and Joy and has two main parts. One is language independent that adjusts some layout details like joining lines and digits. And the second is language dependent.
This latter part principally undoes some pre-processing steps like detokenization, combines sequences of non-ASCII characters into single words, tokenize some punctuation marks and removes non-translated non-En words.

Using a test set consisting of rather short and simple Arabic sentences and 16 references, the system reached a BLEU score of 46.25 after 10 optimization's iterations.
This system was a starting point to build the English to Arabic system presented in the next section.

Chapter IV

English to Arabic system

Based on the previous chapters, building the lecture translator should have acquired a more distinct outline now. Indeed, the task could be reduced to three major measures:

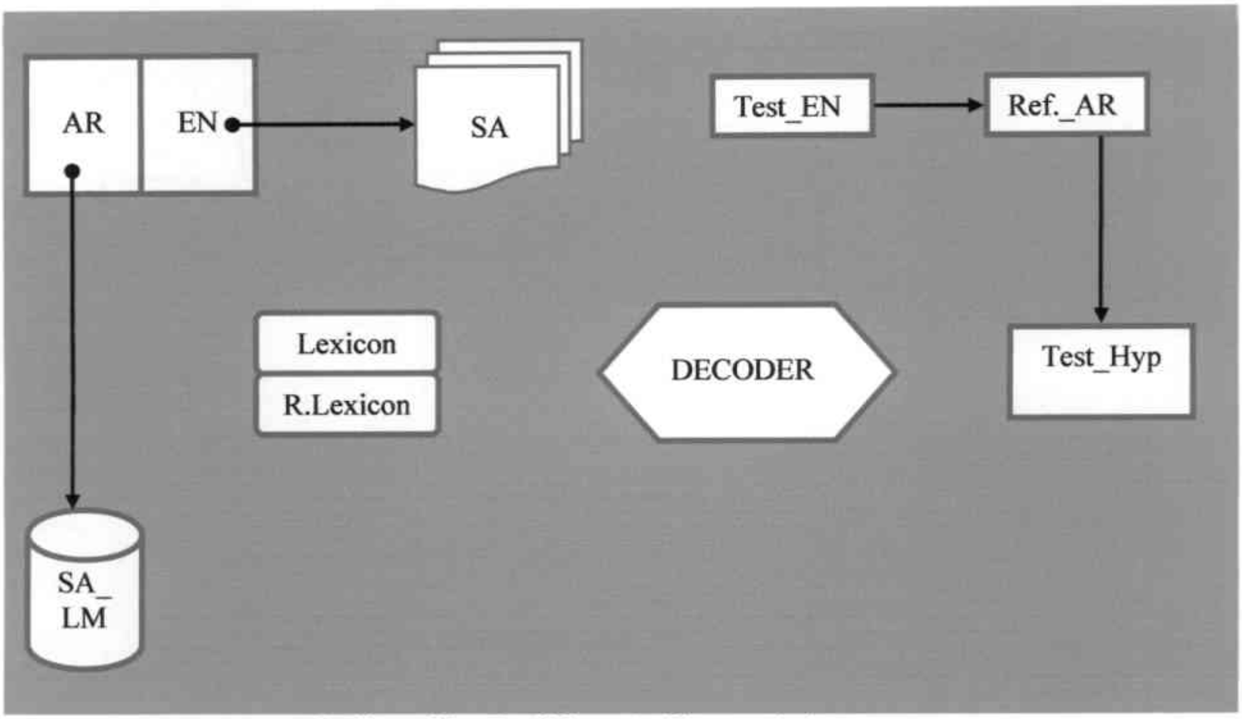
* Building the English to Arabic translation system using the SMT Arabic to English translation system as starting point

* Providing adequate training data, creating test data and generating a new Language Model

* Optimizing the Translator's performance using pre-processed data and multiple language models

4.1 From Arabic-English to English-Arabic:

Let us consider the following schema which offers a simplified outline of the English-to-Arabic translation system

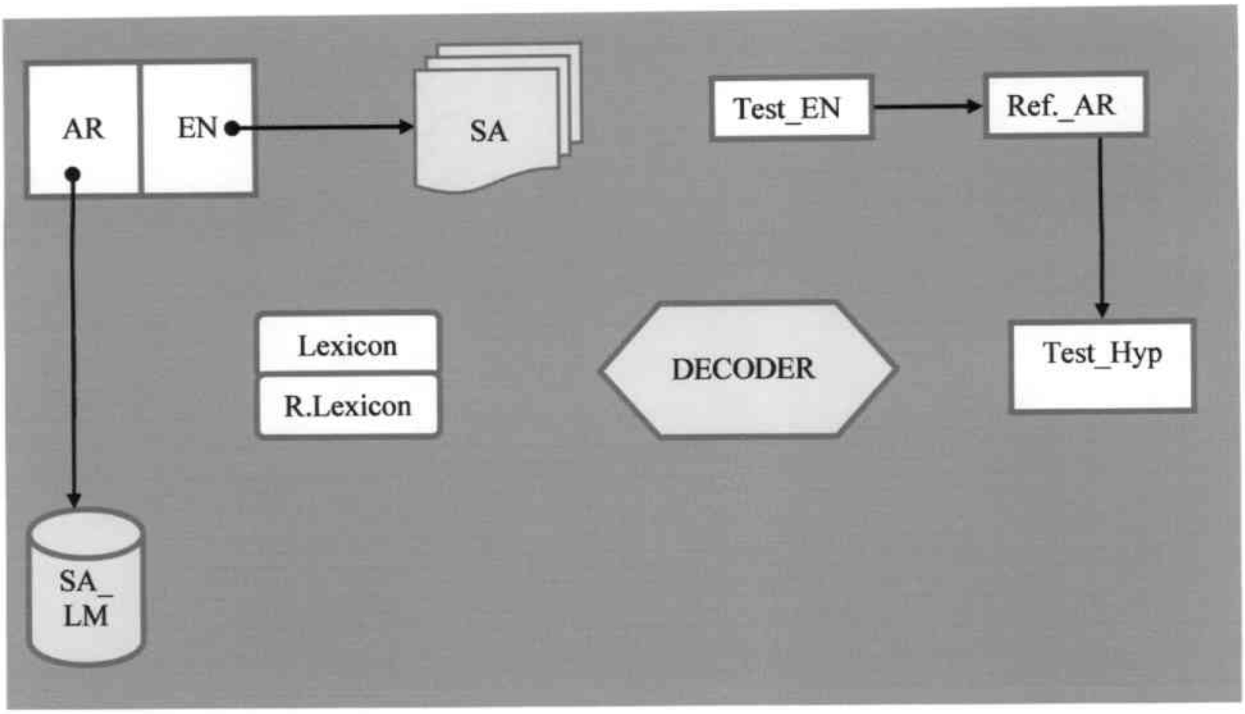


Outline of the English-to-Arabic translation system

The first step described above is to derive this system from the available Arabic-to-English system. Although no new training is required, most of the system's components need to be adjusted.
As shown in the former section, the CMU SMT system outputs both a lexicon and a reverse lexicon while training the Arabic-English bilingual corpus. Inversing the lexica would correspond to training the corpus with inversed source and target languages.

As to the language model, the Arabic side of the training corpus is indexed and used as a Suffix array language model (SALM). Target and source file are inversed and the English side is indexed and given as source file suffix array for phrase extraction.

One more major change concerns the post-processing script. Since the output would be in Arabic, other normalizing and post processing steps are necessary to display the output correctly. In the following schema, the components marked in yellow undergo these adjustments.



First step: Building the English to Arabic translation system using the SMT Arabic to English translation system as starting point

The Translation is done using one of the references of the Arabic-English system as test set and evaluated with the BLEU metric. The score was equivalent to 21.83 which is rather high given that only one unique reference was available, i.e. the former English test set.
The quality of the translation can easily be described as very good by an Arabic native speaker and the unmatched n-grams are still a good translation, though not identical to the reference.

Successfully fulfilling the basic requirements of an English Arabic system meets the first of the three steps listed above. The next sub-section highlights the second one.

4.2 The lecture Translator system

As mentioned in the first chapter, the lecture translator relies on both speech recognition and statistical machine translation to fulfil the required speech to text translation. This makes the task more complicated compared to the introduced translation systems. Unlike text input, spontaneous speech is not expected to be correct and well formed but rather comprising ungrammatical utterances, erroneous phrases and hesitations.

Another major difference is that speech recognition output is delivered without punctuation. This implies a lower BLEU score as it can be noticed below.

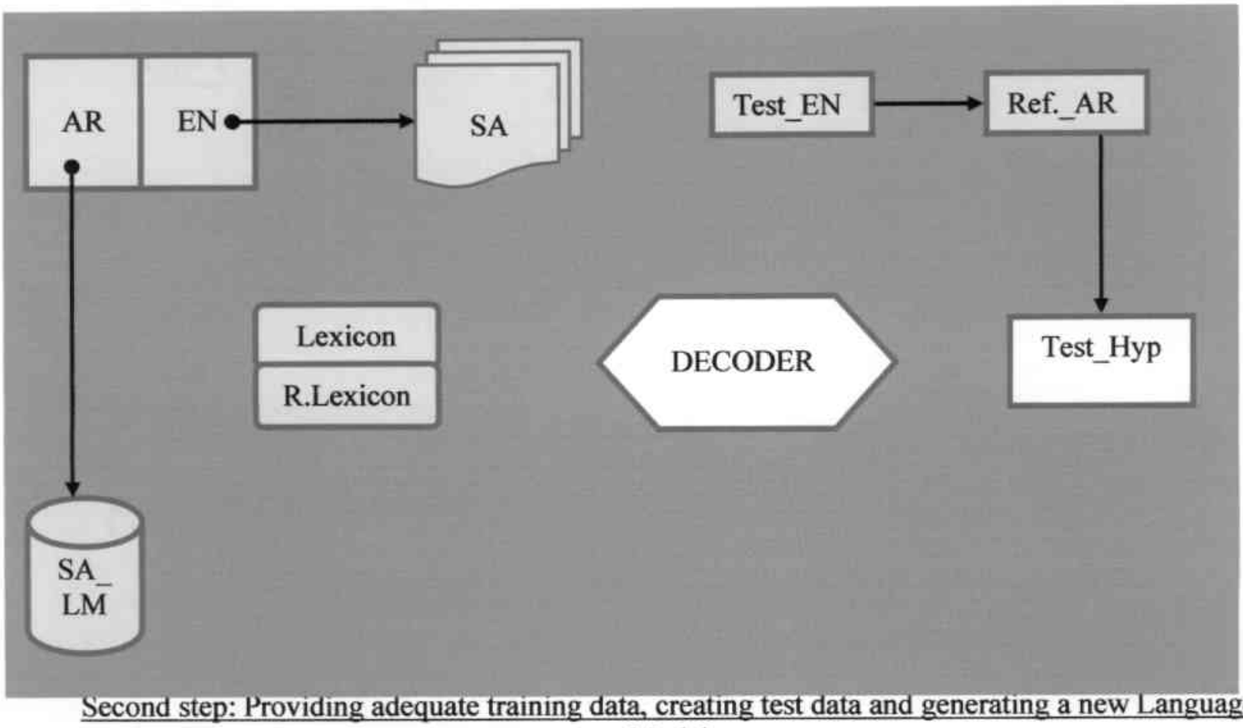
Punctuation not only improves segmentation but it also increases the number of matching n-grams by including the matching punctuation marks. Removing punctuation from the English side of the training corpus and conserving the other decoding parameters, reduced the BLEU score of the English to Arabic translation system introduced above to 14.51.

4.2.1 System components:

As shown before, the available Arabic-English parallel corpus consists principally of news data hardly covering other domains but politics. Yet, the lecture translator designs talks with scientific orientation and especially touching computer science fields like the statistical machine translation and speech recognition.

On another side, this training data often contains interviews and discussions in different confusing Arabic dialects chiefly the Lebanese. The lecture translator, however, is supposed to have an output strictly in modern standard Arabic (MSA).

As a consequence, the translation data's nature strongly limits the translator's performance and is far from embracing the system's required components. The next step is, therefore, to provide the missing data and generate the missing component. Those are again marked in yellow in the following schema.



Second step: Providing adequate training data, creating test data and generating a new Language Model

Creating a test set:

Even if the initial English to Arabic system performed rather well, it obviously would not guarantee the same behaviour if employed for translating lectures. Inevitably, an in-domain test set had to be created along with the corresponding reference. This permits a more realistic, though not totally reliable, evaluation of the eventual translator's quality. The test set corresponded to 100 sentences extracted from a Lecture's transcript. The talk was a spontaneous introduction of the InterACT center and its main projects given in 2005 by Professor Waibel. The data had to be cleaned, i.e. removal of senseless utterances like "uh", "uhm" and then pre-processed.

Generating a reference was a more complicated and a long time consuming task. It had to be done manually with the aid of the BETA version of the Google automatic English to Arabic translation. A second reference was created ulteriorly. The second reference was inspired from the translation's hypothesis in a trial to elucidate the subjective evaluation of the translation's quality.

Generating a language model:

Language models strongly affect the translation system's performance. Usually the more data the better which makes the 120 M words available GALE data sounds rather promising.
Despite its size, the GALE data weakly covers lectures domain. As a consequence, more adequate data needed to be collected to generate an in-domain language model.

The World Wide Web was the unique source and still provided limited results. Compared to other languages like English or Germany etc, the amount of online Arabic data is still modest and mostly not in a process-friendly format (as for instance scanned books outputted in Gif format). A mere example reflecting online Arabic data shortage compared to English would be the number of results for searching for the same word in Arabic and English.
Google search machine outputs more than **47. 5M** results for the input word "Natural Language Processing". While giving the word "معالجـــة اللغـــات الطبيعيـــة" (which means the same in Arabic) delivers no more than **0.26 M** pages written in Arabic and still with no guarantee of pure MSA.

The data was collected using search engines and links posted in forums. It is composed of one translated book about artificial intelligence, many talks about scientific projects, events and meetings, scientific articles with computer science as major theme and different articles about modern information technologies. This reached the size of about 111000 words.
The last steps left were pre-processing the data and generating the suffix array language model as explained in the previous chapter.

Need for new training data

The IWSLT translation data is not sufficient to reach an acceptable lectures translation quality. Since obtaining parallel corpora is even much more complicated than collecting language model data, the available GALE English-Arabic corpus had to provide the necessary training data for the lecture translator's system. Yet the training data size had to be restricted to 5Millions words. Extracting the "best" 5 Millions words from the parallel corpus corresponds to extracting the sentences-pairs containing the maximum n-gram matching with the test set. This is referred to as sub-sampling the training corpus.
More precisely, it consists in outputting all sentences-pairs containing at least one word from the test set in a log file. Each sentence-pair is preceded with an n-grams count. The extracted corpus size can be controlled by limiting the ratio of the n-gram count divided by the sentence length.

Using the first part of the test set, a 4 millions words corpus was sub-sampled and the first experiments could be carried out using the generated language model and 53 sentences from the lecture-test set. The decoding was done offline using a pruned phrase table without punctuation. The BLEU score was 7.22, which could be justified with the effect of the out-of-domain training data as well as the lack of evaluation data.

Further online experiments were carried out in order to observe the effect of the language model nature and size. The results shown in the following table reinforce idea that content of the language model is of a comparable importance to its size.

Language Model		BLEU
Data	Size	
Unseen part of Prof. Waibel Lecture	0.001 millions words	5.6
Collected data	0.09 millions words	6.41
GALE data	Over 100 millions words	7.2

Table 1: Language model effect

Taking a closer look into the translation lattices, some of the common errors could be quickly discerned and directly related to the morphological gap between the two languages. According to previous work, morphological driven pre-processing could reduce this gap and improve the translation quality. This approach has already shown to be effective for many other morphological rich languages like German. It also proved to be helpful in some Arabic-English translation experiments.
In order to improve the lecture translation into Arabic, some pre-processing schemes were applied and compared. They are presented in details in the next subsection.

4.2.2 Morphological analysis and Pre-processing mechanisms

-Definition of morphology
Morphology could be generally defined as the science of form or structure. [15]. In the context of this work it precisely refers to the branch of grammar which studies the structure or forms of words. The main branches are inflectional morphology, derivational morphology, and compounding [16]. Morphology highly differs from one language to another giving place to morphologically very rich languages as well as languages exhibiting rather poor morphology. Arabic and English illustrate these two extremities as explained below.

- Morphological gap between Arabic and English
Inflectional morphology as well as derivational morphology is of an evident potency in the Arabic language. Unlike English, many semantically dependent Arabic words are attached to the next word as a prefix. As examples we cite the definite article "the", some coordinating conjunctions like "and" "then":

Smart and beautiful => ذكي وجميل

Moreover, suffixes are very frequent in the Arabic language. For example, all English possessive pronouns would appear as a suffix attached to the owned “object”.

Their professor => استاذهم

Notice that even more than one suffix or prefix could be attached to the same word (concatenative morphology).

And like the translation => وكالترجمة

Hence, an Arabic word (separated by a white space) usually corresponds to more than one English word. Thus, the majority of the Arabic words are decomposable into prefix, stem and suffix. Many previous attempts have appeared in order to overcome this gap and induce a better morphological symmetry between Arabic and English as trials to improve the statistical machine translation’s performance for this language pair. Based on some of these approaches and pre-processing schemes, we mainly reduce the decomposition schemes adopted in this work to a small set of prefixes. The proposed schemes as well as their effect on statistical machine translation could be observed in the following part.

- Pre-processing mechanisms and results

Finding optimized pre-processing strategies helping improve the translation’s quality between Arabic and English has been interesting many researchers. Many papers have been published proposing various approaches. One of the main common concerns was how to distinguish prefixes and suffixes from real parts of a stem. This task becomes even more complicated when it comes to diacritics-free script. Some proposed solutions are based on frequency, i.e. Deciding whether to split or not would depends on the frequency of the compound noun vs. the resulted stem. Some others put some splitting rules implemented in two automata (one for prefixes and one for suffixes) with prefixes, respectively suffixes, as nodes and an end state corresponding to a existing stem in the text. [17]. Based on previous work, Arabic language knowledge as well as the available data’s nature three pre-processing strategies have been defined to improve the Lecture translator’s performance.

**Adopted strategies and results: As mentioned above, three main splitting schemes have been proposed to improve the lecture translator’s performance. The considered prefixes and suffixes were chosen on the basis of their frequency in the Arabic language generally and in the limited speech domain more particularly. A summarizing table of the involved prefixes and their meanings in the English language looks like the following:

	Group 1		Group2			Group3
Arabic prefixes	و	ف	ب	ك	ل	ال
English translation	And	So/ then	With/ in	like	For / to	The

Separating suffixes is a more complicated task when it comes to avoiding wrong splitting. More complicated strategies are needed to minimize splitting and re-attaching (post-processing step) errors. Thus, and due to the lecture’s nature, only one suffix was taken into consideration. This is done in an attempt to align the frequently-used possessive pronoun “our” to a separated Arabic “word” “نا “.

One further criterion of the choice of splitting strategies relies on some appearance order. This order can be observed in the table above where members of the first group would precede those of the second, which on their turn precede that of the third. This helps detect whether the letters introduced above should be split or not.

The following simple algorithm describes the splitting process:

*Step 1: Split all the “و”, if at the beginning of a word.

*Step 2: Check the letters of the resulting words from left to right

- If a word contains the sequence “ال “ followed by:
  - A letter from Group 2 then white space or
  - A letter from Group 2 then a letter from Group 1 then white space

Then these letters can be considered as prefixes and be split.

*Step3: Split the suffix “نا” when encountered at the end of a word (from left to right)

These steps where partially assigned to three splitting types in order to compare their effects:

Types	Scheme
Type0	Splitting “ و “
Type1	*Type0 + split (ل / ك / ب) if followed with “ ال” in that case split ال too * Separate the suffix „ نا”
Type2	*Split all the „ ف” and „ ال „ prefixes +Type1 steps

The adopted splitting schemes

These types were applied to GALE data, using the first reference of the Gale Arabic-to-English translation as test set. The following table summarizes the obtained results and shows that pre-processing helped improve the translation's quality.

Preprocessing scheme	Stretching Factor	BLEU score
None	~1.2	17.87
Type0	~1.1	18.05
Type1	~1.0	18.03
Type2	~0.9	17.11

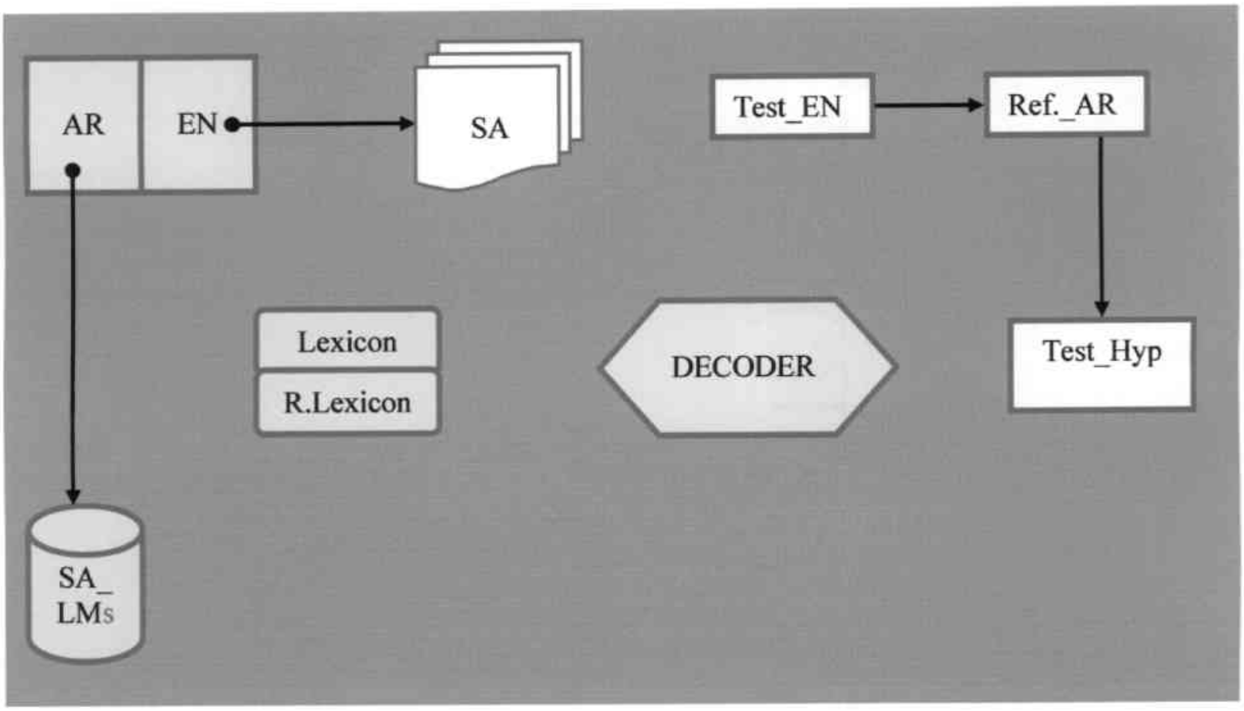
The effect of the different pre-processing types on the translation's quality

Please notice that the split prefixes- respectively suffixes- are attached to the following- respectively preceding- word in the post-processing step. This precedes the evaluation's phase.

**subjective evaluation and choice: The pre-processing schemes were also applied to the Lecture Translator's system. Based on the outputted hypothesis as well as the previous results, the Type1-slightly changed- was decided to be best performing best. The new Type was named Type1.mod. It consists of the same type1 steps provided that the suffix split is abandoned. This is due to the different word order and some distortion model's deficiencies which cause the suffix to be re-attached to the wrong word.

4.2.3 The final Lecture translator System:

The undertaken steps and measures smoothly prepare for the final Lecture Translator System.
This results from gathering and employing all the previous adaptation and optimization's ideas. The system's components concerned with this step are marked in yellow in the following schema:



Step3: Optimizing the Translator's performance using pre-processed data and multiple language models

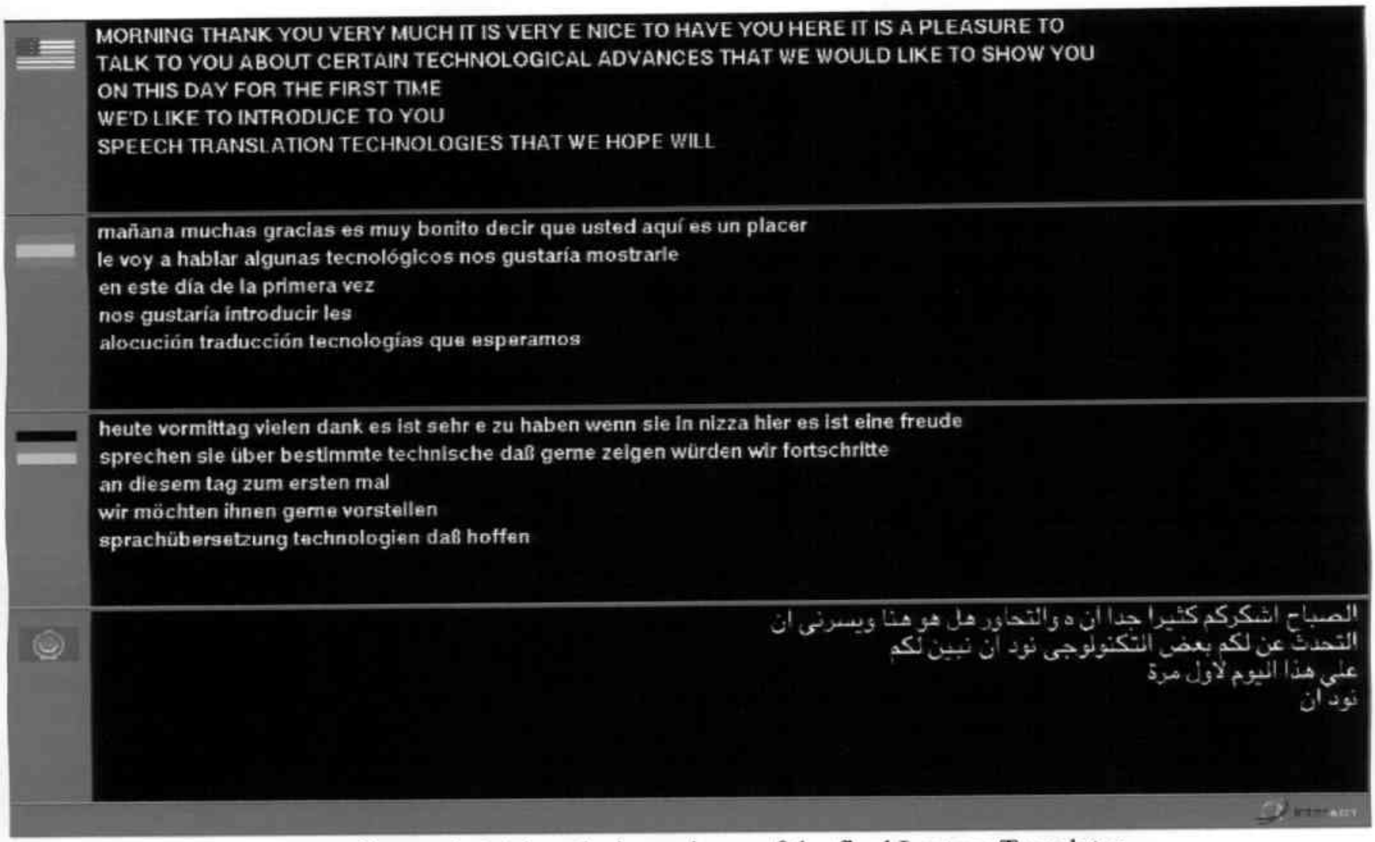
New, more covering training data was sub-sampled using a more covering set of past lectures: all the available past lectures were segmented and cleaned in order to extract new and more flexible training data. Still in the limits of 5Millions words Corpus. The data was pre-processed using the Type1.mod and the whole system was then trained. This outputted a new Lexicon and Reverse Lexicon.
For the final system, two language models were used by the decoder. The effect of the second Language Models could be observed in the following three experiences, where all the parameters, up to the Language models, were kept unchanged.

Experiment	FOU_LM (120M)	Collected_DataLM (0.111M)	BLEU Score
First	X		18.23
Second		X	13.88
Third	X	X	18.77

The effect of Language Model

As a further attempt towards a better performance, both online and offline decoding were combined.

The following screenshot was taken during a demo of the final Lecture Translator:



MORNING THANK YOU VERY MUCH IT IS VERY E NICE TO HAVE YOU HERE IT IS A PLEASURE TO
TALK TO YOU ABOUT CERTAIN TECHNOLOGICAL ADVANCES THAT WE WOULD LIKE TO SHOW YOU
ON THIS DAY FOR THE FIRST TIME
WE'D LIKE TO INTRODUCE TO YOU
SPEECH TRANSLATION TECHNOLOGIES THAT WE HOPE WILL

mañana muchas gracias es muy bonito decir que usted aquí es un placer
le voy a hablar algunas tecnológicos nos gustaría mostrarle
en este día de la primera vez
nos gustaría introducir les
alocución traducción tecnologías que esperamos

heute vormittag vielen dank es ist sehr e zu haben wenn sie in nizza hier es ist eine freude
sprechen sie über bestimmte technische daß gerne zeigen würden wir fortschritte
an diesem tag zum ersten mal
wir möchten ihnen gerne vorstellen
sprachübersetzung technologien daß hoffen

الصباح اشكركم كثيرا جدا ان ه والتحاور هل هو هنا ويسرني ان
التحدث عن لكم بعض التكنولوجى نود ان نبين لكم
على هذا اليوم لاول مرة
نود ان

Screenshot taken during a demo of the final Lecture Translator

Chapter V

Conclusion

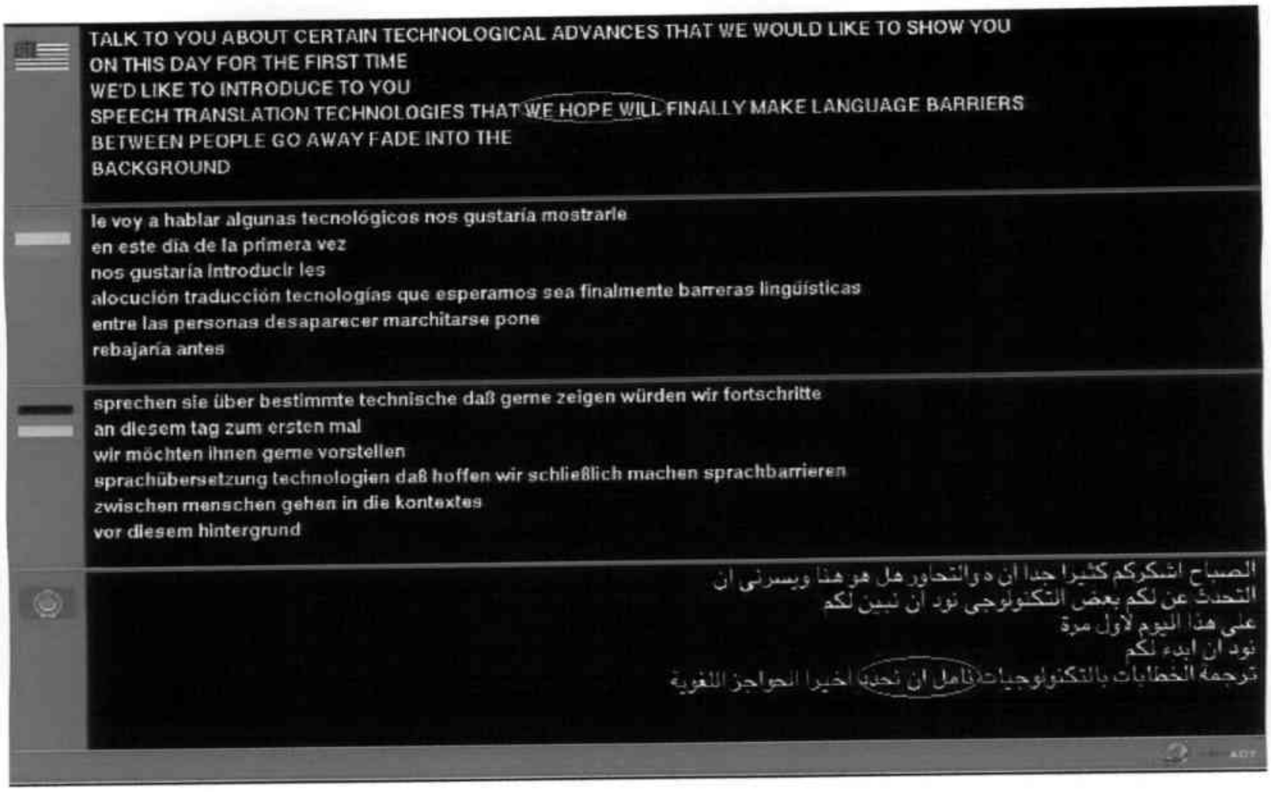
Arabic is a very challenging language to translate to and from. This is due to many reasons such as its long-distance reordering of words, a very rich morphology, script-based constraints, shortage of available Arabic data Many ideas have been proposed to still reach an acceptable translation quality especially in the English-Arabic language pair. In this work we describe the process of expanding the CMU Lecture Translator with an English-to-Arabic statistical translation system. The system was built using IBM out-of-domain training data. Many adaptation's measures had to be taken. This mainly involved pre-processing schemes and the generation of an in-domain language model. During these steps, the following was observed:

- Pre-processing the Arabic data could reduce the morphological gap between English and Arabic. Assumed that the training corpus has a size of about 5 Millions words, splitting some Arabic prefixes improves the translation's quality. It balances the vocabulary size in the parallel corpus and allows a better alignment.

- The language model content has an important impact on the translation's performance. Indeed, adding an in-domain 0.1M words LM to an over than 120M words GALE LM raised the bleu score about 0.5 points.

The final performance, however, proved that these optimisations' efforts are still insufficient. The examples presented below are analysis of some output errors (marked in yellow). This could give a concrete idea about the character of some erroneous translations.

The Lecture Translator should be, thus, a subject to many improvements. These could be concerned with providing more adequate training and test data and dealing with some Arabic specific particularities like the duality problem, pronoun problems
A better pre-processing could help make a step in this direction. Keeping diacritics in the used training data could be experimented. This could offer some improvement, provided that the last letter's diacritics are removed before the training phase or alternatively, removing the diacritics in the Post-processing step.



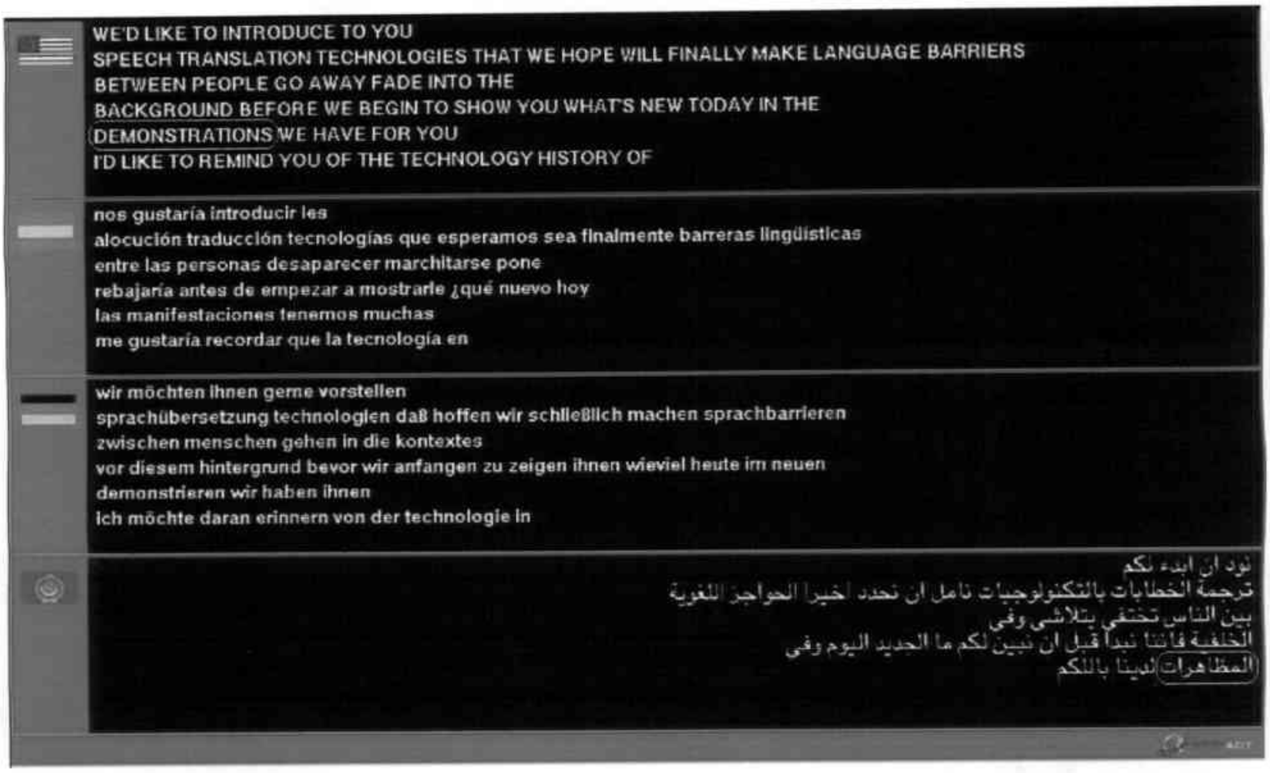
TALK TO YOU ABOUT CERTAIN TECHNOLOGICAL ADVANCES THAT WE WOULD LIKE TO SHOW YOU
ON THIS DAY FOR THE FIRST TIME
WE'D LIKE TO INTRODUCE TO YOU
SPEECH TRANSLATION TECHNOLOGIES THAT WE HOPE WILL FINALLY MAKE LANGUAGE BARRIERS
BETWEEN PEOPLE GO AWAY FADE INTO THE
BACKGROUND

le voy a hablar algunas tecnológicos nos gustaría mostrarle
en este día de la primera vez
nos gustaría introducir les
alocución traducción tecnologías que esperamos sea finalmente barreras lingüísticas
entre las personas desaparecer marchitarse pone
rebajaría antes

sprechen sie über bestimmte technische daß gerne zeigen würden wir fortschritte
an diesem tag zum ersten mal
wir möchten ihnen gerne vorstellen
sprachübersetzung technologien daß hoffen wir schließlich machen sprachbarrieren
zwischen menschen gehen in die kontextes
vor diesem hintergrund

الصباح اشكركم كثيرا جدا ان د والتحاور هل هو هنا ويسرني ان
التحدث عن لكم بعض التكنولوجي نود ان نبين لكم
على هذا اليوم لاول مرة
نود ان ابدء لكم
ترجمة الخطابات بالتكنولوجيات نامل ان نحدد اخيرا الحواجز اللغوية

=> The verb "will make" was related to the pronoun "we" in the Arabic translation. This kind of errors happens often mainly when the distance between the verb and the subject is rather long.



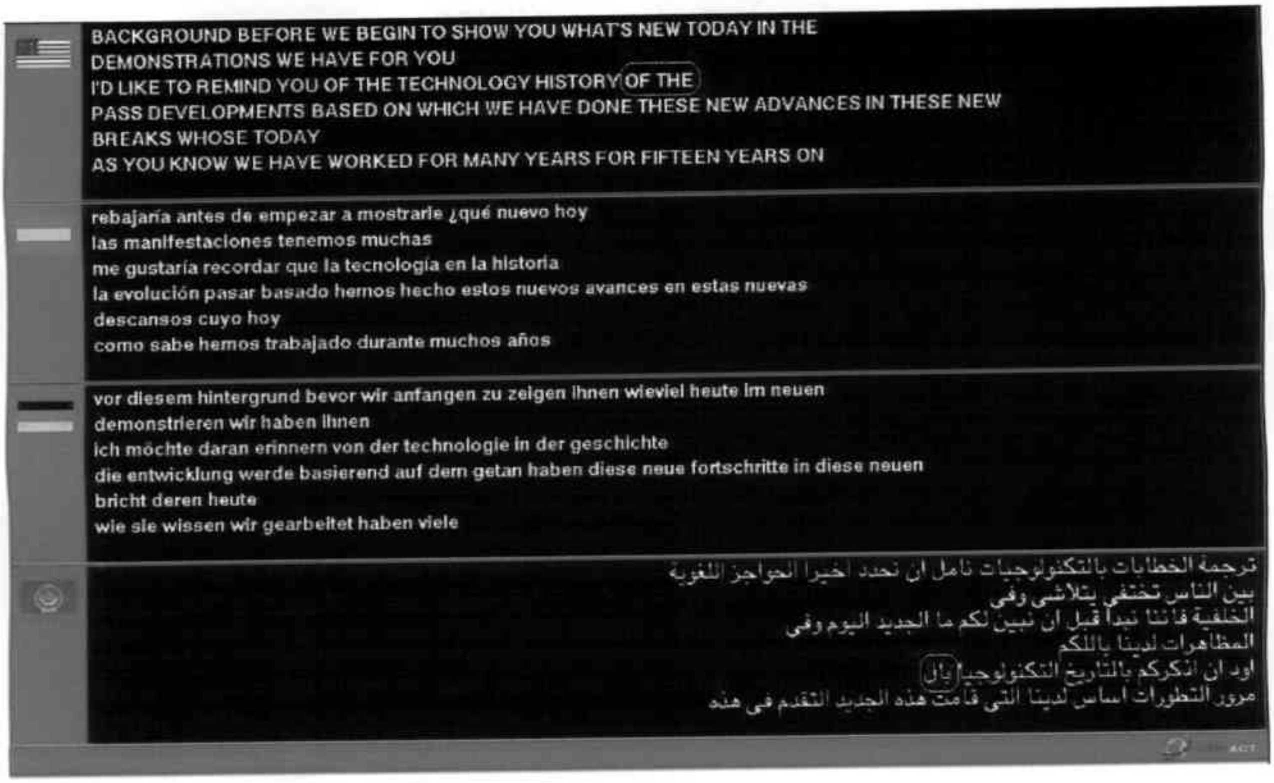
WE'D LIKE TO INTRODUCE TO YOU
SPEECH TRANSLATION TECHNOLOGIES THAT WE HOPE WILL FINALLY MAKE LANGUAGE BARRIERS
BETWEEN PEOPLE GO AWAY FADE INTO THE
BACKGROUND BEFORE WE BEGIN TO SHOW YOU WHAT'S NEW TODAY IN THE
DEMONSTRATIONS WE HAVE FOR YOU
I'D LIKE TO REMIND YOU OF THE TECHNOLOGY HISTORY OF

nos gustaría introducir les
alocución traducción tecnologías que esperamos sea finalmente barreras lingüísticas
entre las personas desaparecer marchitarse pone
rebajaría antes de empezar a mostrarle ¿qué nuevo hoy
las manifestaciones tenemos muchas
me gustaría recordar que la tecnología en

wir möchten ihnen gerne vorstellen
sprachübersetzung technologien daß hoffen wir schließlich machen sprachbarrieren
zwischen menschen gehen in die kontextes
vor diesem hintergrund bevor wir anfangen zu zeigen ihnen wieviel heute im neuen
demonstrieren wir haben ihnen
ich möchte daran erinnern von der technologie in

نود ان ابدء لكم
ترجمة الخطابات بالتكنولوجيات نامل ان نحدد اخيرا الحواجز اللغوية
بين الناس تختفي بتلاشى وفي
الخلفية فاننا نبدأ قبل ان نبين لكم ما الجديد اليوم وفي
المظاهرات لدينا باللكم

=> This screenshot illustrates a good example of a false translation of the word "demonstrations" which is due to the out-of-domain training data's nature. Indeed the word "demonstration" was translated in the meaning of "manifestation" rather than "presentation".



BACKGROUND BEFORE WE BEGIN TO SHOW YOU WHAT'S NEW TODAY IN THE
DEMONSTRATIONS WE HAVE FOR YOU
I'D LIKE TO REMIND YOU OF THE TECHNOLOGY HISTORY OF THE
PASS DEVELOPMENTS BASED ON WHICH WE HAVE DONE THESE NEW ADVANCES IN THESE NEW
BREAKS WHOSE TODAY
AS YOU KNOW WE HAVE WORKED FOR MANY YEARS FOR FIFTEEN YEARS ON

rebajaría antes de empezar a mostrarle ¿qué nuevo hoy
las manifestaciones tenemos muchas
me gustaría recordar que la tecnología en la historia
la evolución pasar basado hemos hecho estos nuevos avances en estas nuevas
descansos cuyo hoy
como sabe hemos trabajado durante muchos años

vor diesem hintergrund bevor wir anfangen zu zeigen ihnen wieviel heute im neuen
demonstrieren wir haben ihnen
ich möchte daran erinnern von der technologie in der geschichte
die entwicklung werde basierend auf dem getan haben diese neue fortschritte in diese neuen
bricht deren heute
wie sie wissen wir gearbeitet haben viele

ترجمة الخطابات بالتكنولوجيات نامل ان نحدد اخيرا الحواجز اللغوية
بين الناس تختفي يتلاشى وفى
الخلفية فاننا نبدأ قبل ان نبين لكم ما الجديد اليوم وفى
المظاهرات لدينا باللكم
اود ان اذكركم بالتاريخ التكنولوجيا بال
مرور التطورات اساس لدينا التى قامت هذه الجديد التقدم فى هذه

⇨ Here we can discern the impact of the applied pre-processing schemes. The words "of the" would not have been translated correctly otherwise. Indeed, they would appear as prefix in the Arabic output so no correct equivalent would have been outputted. The only deficiency- which is due to bad speech segmentation here- is that the translation should have been attached to the next Arabic word to get a perfect result.

Bibliography

[1] August 23, 2006: How to Talk Like an Iraqi - Laptop software that can translate English-Arabic conversations on the fly is being tested in Iraq. By Kate Greene. Technology Review (August 23, 2006).

[2] http://www.myke.com/enigma.htm

[3] www.ics.mq.edu.au/~szwarts/Research.php

[4] http://www.thocp.net/reference/machinetranslation/machinetranslation.html

[5] http://www.is.cs.cmu.edu/

[6] ESSLLI Summer Course on SMT (2005) by Chris Callison-Burch and Philipp Koehn

[7] Daniel Jurafsky & James H.Martin, Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition, August 3, 2006

[8] A statistical MT tutorial workbook, Kevin knight, 1999

[9] Sadat, Fatiha & Nizar Habash. Combination of Arabic Preprocessing Schemes for statistical Machine Translation

[10] http://www.tcstar.org/pubblicazioni/scientific_publications/Elda/tcstar_lrec06.pdf)

[11] Papineni et al., 2001

[12] http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/phrase2003.pdf

[13] The CMU Statistical Machine Translation System for IWSLT 2005
Sanjika Hewavitharana, Bing Zhao, Almut Silja Hildebrand, Matthias Eck,
Chiori Hori, Stephan Vogel and Alex Waibel

[14] http://www.darpa.mil/ipto/programs/gale/

[15] www.sdvc.uwyo.edu/grasshopper/ghnmglos.htm

[16] www.essex.ac.uk/linguistics/clmt/MTbook/HTML/node98.html

[17] Morpho-syntactic Arabic Preprocessing for Arabic-to-English Statistical
Machine Translation [Anas El Isbihani, Shahram Khadivi, Oliver Bender, Hermann Ney]