Kandidatuppsats

Machine Translation, universal languages and Descartes

MT systems not always produce accurate translation results. This study investigates the nature of these problems and examines the benefits of universal languages to improve the quality for these systems.

Bamdad. Payvar
Institute of technology
Blekinge Tekniska Högskolan
bamdapidayvar@msn.com

Handledare: Stefan.johansson

ISSN-nummer:
Abstract

The aim of this thesis is to explore Machine Translation and the problems that these system are experiencing when translation between two different languages. The grammatical structures will be studied for English, Swedish and Persian to find a common pattern that could relate different ideas in each language to each other.

In the other hand an inter lingual MT will be developed according to “René Descartes” principals that not only produces translations to English, Persian and Swedish but it even provides a new way of inputting text just by clicking buttons which each represent a word or concept. Then the system will be presented to a group of chosen users to study the human interaction with the application and identifying new problems associated with the new developed system and evaluating the results.

The specific objectives are: the role of prepositions and other grammatical structures in determining the meaning of a text. The study even examines the possibility of using Descartes theory for improving Machine Translation.

The study was conducted in “BTH”. The data was collected through research, experiments, and Self-reporting.
CONTENTS
1. Introduction

1.1 Background

2. History of Machine Translation
It is possible to trace ideas about MT back to seventeenth century in connection with ideas on universal languages by Descartes, Leibniz, Wilkins, etc. in the mid 1930s Georges Artsrouni and Petr Trojanski applied for patents for their translating machines. “Trojanskijs model proposed not only a bilingual dictionary but also a scheme for coding interlingual grammatical roles based on Esperanto. By 50’s Andrew Booth and Warren Weaver put forward the first tentative ideas for using the newly invented computers for translating natural languages. The Georgetown experiment (1954) involved fully automatic translation of over sixty Russian sentences into English. The experiment was a great success and ushered in an era of substantial funding for machine-translation research. The authors claimed that within three to five years, machine translation would be a solved problem. In the same year the MIT group led by Yngve received funding from the national science foundation after seeing a demonstration of the IBM-Georgetown system.” (Hutchins) Since then a great deal of work has improved the quality of the translations. Today there are a few well-known names such as Systran, Google, Bing, Babel fish…etc working on this topic.

3. Problems associate with Machine Translation
It is no secret that the MT systems have always suffered from difficulties in producing trustworthy results. The problem is so bad that it is widely common to accept an understandable translation as ‘good enough’ for less important purposes regardless of eventual grammatical or structural mistakes. (Translation and the structure of language – Sydney M. Lamb)
Sydney M.Lamb wrote: “…what impressed me more than anything else when I got into MT research and looked around at what other researchers were doing was that most of them had little or no training or experience in linguistics, and they didn’t seem to think that such training and experience was necessary for their work.” (Translation and the structure of language- Sydney M.Lamb)

Thinking rationally one would sooner or later come to this very important conclusion that there would be No accurate translation without a correct interpretation of the source. This concept is described by H.L.Capron and J.A.Johnson as following:

“Some natural languages words are easy to understand because they represent a definite item like ‘table’, ‘car’ and ‘food’ for example. Some other words however are much too abstract to lend themselves to straightforward definitions: ‘justice’, ‘virtue’ and ‘beauty’, for example. But this kind of abstraction is just the beginning of the difficulty. Consider the word “hand” in these statements:

- Morgan had a hand in the robbery.
- Morgan had a hand in the cookie jar.
- Morgan is an old hand at chess
- Morgan gave Sean a hand with his luggage.
- Morgan asked Marcia for her hand in marriage

As you can see natural languages abounds with ambiguities; the word ‘hand’ has a different meaning in each statements. In contrast sometimes statements that appear to be different really mean the same thing: ‘Alan sold Jim a book for five dollars’ is equivalent to ‘Jim gave Alan five dollars in exchange for a book.’

Feeding computers the vocabulary and grammatical rules they need to know is a step in the right direction. However, as you saw earlier in the account of the language translation fiasco, true understanding requires more: Words must be taken in context. Humans start acquiring a contest for words from the day they are born. Consider this statement: Jack cried when Alice said she loved Bill. From your own context, several possible conclusions can be drawn: Jack is sad, Jack probably loves Alice, Jack probably thinks Alice doesn’t love him, and so on. These conclusions might not be correct but they are reasonable interpretations based on the contest the reader supplies. On the other hand, it would not be reasonable to conclude from the statement that Jack is a carpenter or that Alice has a new refrigerator. (“(Tools for an information age, H.L.Capron-J.A.Johnson/548)
4. **Natural Languages and variations of worldview among nations**

Every nation has a special historical progress and unique experiences. These nations then build their own culture and worldview upon these experiences. The worldview of a nation reflects on its language. (J-Allwood & L-G Andersson/30)

This explains the fact that different cultures and nations use different ways for expressing ‘elements’ in their language. For instance in Persian if they want to refer to:

- ‘Alice lurking ted’ they would say: ‘Alice put a hat on teds head’.

The meaning of these Phrases is the same as you can see even if they are expressed in totally different manner using different words and elements. This is the very essence of the idea behind ‘Universal Languages’ as a tool to mediate ideas among different languages.

5. **Universal Languages**

6. **Universal Grammatic**

Sometime during 16´th century gave a group of rationalist logics, semantics and grammarians of Port-Royal convent in France, two works by combining medieval speculative grammar with Descartes' philosophy advocated a kind of universal grammar. the group's leading members were the two Frenchmen Arnauld and Lancelot. They wanted to create:

a general or universal grammar valid for all languages

Rational or rationally-based grammar, based on characteristics of the human mind.

Port Royal grammar argued that any linguistic expression had an inner mental side and an outer physical, a sequence of sounds or written symbols. The mental structure for a phrase is common to all people and thus the basis for the universal grammar. Every ‘situation’ is composed of a set of simple judgments that through these rules (movements of the soul) are assembled into a complex review. (J-Allwood & L-G Andersson/24)

7. **The idea of an ideal language**

Rationalists wanted to create a logical language that was in accordance with the human intellect and nature. Because human intelligence is common to all people, this language could be used by everybody to communicate with each other. the German philosopher Leibniz was a huge supporter of this idea, saying that the universal language would be a kind of calculus that made it easier to reason logically. (J-Allwood & L-G Andersson/26)
8. Descartes Discusses the Idea of an Artificial Language

In a letter to theologian, philosopher, and mathematician Marin Mersenne, philosopher, mathematician and physicist René Descartes proposed an artificial universal language, with equivalent ideas in different tongues sharing one symbol.

"The notion of a universal language was based upon the idea of precisely cataloguing the elements of the human imagination. The great advantage of such a language would be that it would represent everything distinctive”.


However the infinite variations of human imagination seem to make it if not impossible but extremely hard to identify and classify all ‘movements in human soul’ but it seems to be possible and practical to start the process of identifying and categorizing with the most common in use ‘elements’ such as:

- The sense of appreciation
- The sense of danger
- The idea of being careful

For instance consider following phrase: “look out”. This phrase according to the Descartes idea could be identified and classified as “The idea of being careful”. When translating to other languages this categorization would help MT system to understand the meaning of this phrase and therefore finding equivalents in the target language would be much more accurate.
8.1 The purpose and the aim
In many cases the result of sentences that translates to other languages suffers from ambiguity and disorder when using common Machine Translation services.

- The purpose of this study is to investigate the reasons behind this obstacle and identify the problems associated with it. Finding out about previous solutions and theories which claimed to be the ‘answer to the question!’ and in particular paying more attention to ‘Universal Languages’ and

- The aim of the study is to determine how an implementation of Descartes theory would look like and whether this system would be able to produce trustworthy translations.

8.2 Research question
The research is meant to find answers to these fundamental questions:

- What are the reasons behind ambiguity and disorder in translation results when using common MT services?

- Rene Descartes suggested that using an interlingual dictionary was the way to an accurate translation. Would such solution lead to improvement of MT technology and deal with problems associating with translation ambiguities?

8.3 Methodology
This study is done in following steps:

9. Investigation
To be able to study a problem one should understand the nature of the problem first. Choosing “Google” and “Systran” as the projects means of translation service, a set of experiments are required to identify the eventual problems with the translation and the elements associated with them. The experiments are based on 3 languages: Persian, English and Swedish. The results suggest something interesting regarding verbs. It seems that each interpretation of a verb in Persian could be identified to one group. Each group is then influence the syntax of the sentence in its own characteristics.

10. the role of prepositions in determining the meaning of a string
If we consider the Persian verb: “Jangidan” that means “to fight” the following variation could be exist with “John” and “Karl” as the subject and object of the sentence:
- John ba Karl mijangad. (John fights Karl)
- John baraye Karl mijangad (John fights for Karl)
- John hamdoshe Karl mijangad (John fights by Karl’s side)

11. Some other interesting results
Another interesting fact about Persian is that it is possible to build a verb out of a substantive or even adjective. For example “Ziad” is an adjective and means “much” or “a lot”. However “Ziad kardan” is a verb and means “to increase”. Another example is “Asfalt” that is a substantive and refers to “Asphalt” as a noun but “Asfalt kardan” is a verb and means: “To put Asphalt on a surface”. I notice that this is a very important issue and identify it as one of the factors that affects results when translating to Persian. If not handled according to special rules this issue would certainly cause a bad result.

12. How could Descartes idea be implemented in an application
One interpretation of Descartes idea consists of a set of database tables. One for each language for example Swedish, English and so on. Each row represents a unique expression and stores a set of words and data such as preposition and identifiers. In this way each row in a table is related to its equivalent in other tables by the shared symbol or foreign key in database systems.

As you can see in the illustration by relating different concept in different languages with a universal identifier, for example a unique integer it would be easy to determine the true meaning of
a phrase like “look out” in other languages supported by the system based on finding the shared id which in this case is ‘123’.

13. Implementation

14. Integration of Descartes theory and implementation in database systems
Part of the experiments needs an environment to perform. This environment consists of a web-based application which allows the user interacting with the system on the internet. A set of database tables are created to simulate the ‘inter lingual Dictionary’ mentioned by Descartes. Each row of this table consists not only of a word but even additional Meta data relating to the word. Some examples of these Meta data are properties such as the type of the word (verb/substantive/adverb…), the gender if exist (male/female) or prepositions etc. and of course the universal unique “id” which acts as the “sharing symbol” among different db tables.

![Database Table Example](image)

Figur 3. some of the meta data presented along with the word in each row of the database table

15. Experiment environment, implementin operators
There are many operators that decide the form of the sentences and their meaning.

- (?) Usually at the end of a sentence defines the sentence as a question.

- (what/where/who/how…etc) usually at the beginning of a sentence also define the sentence as a question but with a special purpose.

- (past/present) defines the time format for the sentence
These and some more similar operations are implemented as attributes. When processing a sentence the logic of the application would know how to put the words together simply by consulting the value of these attributes along with a set of grammatical rules.

16. Evaluation
Having the experimental environment ready it is time to perform the experiment itself. The purpose of the experiment is to study how human users interact with the system, what are the results and how the users evaluate the system and what is their opinion. In this experiment the user is simply asked to describe a picture by clicking buttons. The MT then builds sentences and paragraphs in Persian, English and Swedish. The quality of the sentences and whether the translation is correct or not will be expressed and registered by the user along with some more data. This data helps to understand if the project has come any closer to achieve its goals or not.

2. Experiment design

There are two separate experiments that are done. In the first experiment by using Google’s online translating service, a set of phrases will be translated from English to Persian and Swedish. The results are then become evaluated. In the second part the same phrases are being produced by Descartes integrated MT and then the quality of the results are evaluated. Beside that the Test persons have an opportunity to describe their opinions about the system. To be able to evaluate the “Descartian’s MT” system a simple test is arranged. A web based environment is prepared. The purpose of the experiment is to see how normal people interact with the system by clicking buttons. As a bonus the translation of the input in Persian, English and Swedish produced by the test person is stored in a data base along with some other input data regarding the test. The test persons are selected carefully and are few. Each test person is representing a nationality, gender, age group and the result of the experiment is stored in the database for future analyses. The test persons are introduced to the system some time ahead so they are familiar with the system and the functions. There is even a user manual created and presented at sight to assist the users.

2.1 Experiment scenario
-The user starts the application at: http://www.student.bth.se/~bapa03/mt/

-The user chose his/her favourite input language presented in the panel.
-User comes to a new webpage with buttons and a phrase in English. The task is to describe the recreate the phrase by clicking the buttons. When done the user clicks on the “submit test” button. and the second of totally four phrases is appeared.

-The user evaluates the quality of the results and writes his/her evaluation of the system in a form. There are some more questions that will be “checked” by the test person. These questions are:

- Sex {male / female}
- What is your First Language? {Swedish, English, Persian, others}
- What age category you belong? Please choose one of the following options: {0-20, 20-40, 40-60- 60-100}
- Was it easy to work with the application? {Easy, medium, hard, impossible to understand}
- How was the quality of the sentences/ result in your opinion? {Poor, middle, good, very good}

2.2 Experiment boundaries
Considering the aim of this study the Descartes inspired MT system is only equipped with minimum amount of vocabulary and grammatical rule sets. However all means are provided to build following sentences in all 3 target languages:

- Morgan had a hand in the robbery.
- They have a hand in the cookie jar with the money.
- Morgan is an old hand in chess.
- 4-I give a hand to Sean.
3. Results

Here are the results for the first phase of the experiment using Google translation online service.

1- Morgan had a hand in the robbery.
   • Morgan hade ett finger med i rånet.
   • مورگان دست در سرقت

2- Morgan had a hand in the cookie jar.
   • Morgan hade en hand i kakburken.
   • مورگان دست در شیشه کوکی برای نگهداری اطلاعات استفاده می

3- Morgan is an old hand at chess
   • Morgan är en gammal hand i schack
   • مورگان دست قدیمی در شطرنج

4- Morgan gave Sean a hand with his luggage.
   • Morgan gav Sean en hand med hans resgods
   • مورگان داد شان دست با توشه خود را

Here are the results for the next phase of the experiment performed by Test users:

Test person 1:
Sex: female
Age: 20-40
First language: Japanese
Finds working with the system as: hard
Evaluates the quality as: good
Feedback:
I have never tried such kinds of translation program before. Therefore, i found it difficult to engage myself in, which is my personal opinion. There are lots of options and buttons users can command, which is clear and simple. When I am typing the technological problems happened due to the BTH website regulations, which was not so positive for me... X( because i dont use programs and computers in general often, so all the troubles got me confused. I like the idea that I can see the words in Persian and Swedish though i know them in English. They are so helpful and also interesting!!

1-Morgan had a hand in the robbery.
   - Morgan var delaktig i rånet.
   - Morgan dar serghat dast dasht.

2-they have a hand in the cookie jar with the money.
   - de fiflar med pengarna.
   - anha dar omore eghtesadi khalaf mikonand.

3-Morgan is an old hand in chess.
   - Morgan är bra på chess.
   - Morgan dar shatranj khebre mibashad.

4-I give a hand to Sean.
   - jag hjälper Sean.
   - Man be Sean komak mikonam.

Test person 2:

Sex: Male
Age: 20-40
First language: Persian
Finds working with the system as: medium
Evaluates the quality as: good

Feedback:
I think it’s better to have a way to go back to the previous question in the test. There are some grammatical mistakes seen in the results but the translation to Persian has good quality. I think this is a good method for using in the schools for teaching foreign languages.
1-Morgan had a hand in the robbery.
   - Morgan var delaktig i rånet.
   - Morgan dar serghat dast dasht.

2-they have a hand in the cookie jar with the money.
   - de fiflar med pengarna.
   - anha dar omore eghtesadi khalaf mikonand.

3-Morgan is an old hand in chess.
   - Morgan är bra på chess.
   - Morgan dar shatranj khebre mibashad.

4-I give a hand to Sean.
   - jag hjälper Sean.
   - Man be Sean komak mikonam.

Test person 3:

Sex: Male
Age: 20-40
First language: German
Finds working with the system as: hard
Evaluates the quality as: middle

Feedback:
Understanding the task was a bit difficult, without reading the manual it is even more hard. Having an introduction could be helpful.

1-Morgan had a hand in the robbery.
   - Morgan var delaktig i rånet.
   - Morgan dar serghat dast dasht.
2-they have a hand in the cookie jar with the money.

- de fiflar med pengarna.
- anha dar omore eghtesadi khalaf mikonand.

3-Morgan is an old hand in chess.

- Morgan spelar med chess.
- Morgan ba shatranj bazi mikonad

4-I give a hand to Sean.

- jag hjälper Sean.
- Man be Sean komak mikonam.

Test person 4:

Sex: Male
Age: 20-40
First language: Persian
Finds working with the system as: hard
Evaluates the quality as: good

Feedback:
I think that the user manual in this site is good but it could be better and more helpful to help users for first try, I think its interesting to work with this software and if the options be more it would be more interesting and at last i should say that there is some problems in some verbs inflection in Persian.

1-Morgan had a hand in the robbery.

- Morgan var delaktig i rånet.
- Morgan dar serghat dast dasht.

2-they have a hand in the cookie jar with the money.

- De fiflar med pengarna.
• anha dar omore eghtesadi khalaf mikonand.

3-Morgan is an old hand in chess.
• Morgan spelar med chess.
• Morgan ba shatranj bazi mikonad

4-I give a hand to Sean.
• Jag hjälper Sean.
• Man be Sean komak mikonam.

Test person 5:
Sex: Female
Age: less than 20
First language: Swedish
Finds working with the system as: impossible
Evaluates the quality as: middle
Feedback:
Jag har aldrig testat på något sånt här program förens nu, det var både svårt men ändå roligt när man fattade hur man skulle göra tillslut. alltid kul och lära sig något nytt :)

1-Morgan had a hand in the robbery.
• Morgan var delaktig i rånet.
• Morgan dar serghat dast dasht.

2-they have a hand in the cookie jar with the money.
• de fiflar med pengarna.
• anha dar omore eghtesadi khalaf mikonand.

3-Morgan is an old hand in chess.
• Morgan spelar med chess.
Test person 6:

Sex: Male
Age: 20-40
First language: Persian
Finds working with the system as: medium
Evaluates the quality as: very good

Feedback:
That was so fascinating and fun, and working with that is not hard specially after reading the user manual it became more easier. but i think that the environment could be more fun and interesting and may be adding some sounds when you are working with tabs change its environment, even i think if it could pronounce the sentences it had been more usefull and attractive

1-Morgan had a hand in the robbery.
   • Morgan var delaktig i rånet.
   • Morgan dar serghat dast dasht.

2-they have a hand in the cookie jar with the money.
   • de fiflar med pengarna.
   • anha dar omore eghtesadi khalaf mikonand.

3-Morgan is an old hand in chess.
   • Morgan spelar med chess.
   • Morgan ba shatranj bazi mikonad

4-I give a hand to Sean.
• jag hjälper Sean.

• Man be Sean komak mikonam.

Test person 7:

Sex: Male
Age: over 60
First language: Swedish

Feedback:

• jag har testat ditt program, men måste erkänna, att vid min höga ålder och med den här typen av program, är det komplicerat att förstå hur man går till väga. Eftersom jag inte kan kolla i manualen hur man går till väga, är jag rädd att jag måste ge upp, eftersom jag inte kan komma vidare och skapa korrepta meningar på engelska.

4. Discussion

The first phase of the experiment shows that even if the results are grammatically correct but they fail to mediate the true meaning of the source. It is clear that each of which of the statements in the experiment are referring to a different situation and the word ‘hand’ has different meaning in each situation. Although the Swedish results seems to be grammatically correct but it fails to link to the true meaning behind the words. It is only one case that the result is correct in both semantic and Grammatik:

Morgan had a hand in the robbery ➔ Morgan hade ett finger med i rånet.

The Persian results are disastrous and in many of the cases Seem to be just a random combination of ‘Morgan’, ‘hand’ and other elements in the source. What ever technique that is used it is not obviously working for all languages well.

By looking at the results of the second phase of the experiment and compare them to the results produced by Google there is a major difference in quality.

-Morgan had a hand in the robbery ➔ Morgan var delaktig i rånet.
Especially when we compare the Persian results with the result from Google even though the source had been exactly the same but the difference is huge.

The Descartes inspired MT has been successful to produce good and accurate results both in Persian and Swedish and therefore confirms the hypotheses “the use of Descartes principles leads to improvement of MT technology and deals with problems associating with translation ambiguities”.

However it should be highlighted that the two systems used in the experiments are fundamentally different in performance and use. There are of course some problems associated with the system yet. While some of the test persons start interacting with the system at once, many users were experiencing difficulties with the system and complain that they were not really sure how to interact with the system. Some were sceptical because of some grammatical mistakes in English results however the goal of this study is mainly focusing on mediating the true meaning and nothing else.
5. Conclusions

The study shows that it is important for any MT system to know the meaning of the source in order to produce trustworthy results. Natural languages are sometimes very complex and they not always logical. Therefore any attempt to translate a sentence by translating its components and applying them to the grammatical rules and constraints will result in ambiguous and wrong results.

The MT model used in the experiments shows that any system which takes the benefit of combining the abilities of an ‘Interlingual Dictionaries’ along with a set of logic and constraints will be able to enjoy the accuracy of the results.

This study suggests that collaboration among linguists and computer scientist is needed to identify all/most of the ‘elements of human imagination’ and register them in a set of database tables to build an Interlingual dictionary. This database would be an asset to MT systems who chose to apply Descartes principals discussed before and will help them provide trustworthy translation among nations.
6. Future perspectives

During the project a new concept was born. The concept of inputting text by clicking buttons rather than spelling each word by its alphabet. To study this technique and developing smart interfaces for mobiles and other gadgets would be an interesting challenge with maybe creative results.

In Google it is possible to copy and paste a text for having a translation. In this system not! (The system that was used for experiments). Future studies are required to find a way to provide the same ability for this system because it does not offer the full range of services needed for ‘everyday tasks’ as it operates now.
7. References


- Batchelor, The Republic of Codes: Cryptographic Theory and Scientific Networks in the Seventeenth Century, 1999
