



UNIVERSIDAD POLITÉCNICA DE CARTAGENA

HUMAN EVALUATION OF MT
-PROYECTO FINAL DE CARRERA-

Autor: Ginés Mendoza Mompeán
Director: Fernando Daniel Quesada Pereira

Cartagena, Septiembre 2013

TABLE OF CONTENTS.....

1. INTRODUCTION	1
1.1 ABSTRACT	1
1.2 OBJECTIVES	1
2. MACHINE TRANSLATION	2
2.1. INTRODUCTION TO MACHINE TRANSLATION	2
2.1.1. LIMITATIONS	3
2.1.2. DIFFERENT KINDS OF MACHINES TRANSLATION	3
2.2. EVALUATION	5
2.2.1 MEASURE “DARPA” OF EVALUATION FOR THE TRANSLATION	5
2.2.2 AUTOMATIC EVALUATION	6
2.2.3 HUMAN EVALUATION	8
3. DEVELOPMENT SYSTEM	9
3.1. DEVELOPMENT TOOLS	9
3.1.1 JAVA SE	9
3.1.2 DEVELOPMENT ENVIRONMENT	10
4. TOOLS FOR HUMAN EVALUATION OF MT	10
4.1. HOW TO USE THE APLICATION FOR HUMAN EVALUATION OF MT	11
4.1.1 HOW TO OPEN THE FILES	11
4.1.2 PARTS OF THE APPLICATION	12
4.1.3 HOW TO SAVE THE EVALUATION OF THE ANNOTATOR	14
4.1.4 RESULT (EXPLICATION)	14
4.1.4.1 TOOLS TO EVALUATE THE TRANSLATION	16
5. THE RESULTS FROM THE EVALUATION OF THE TRANSLATIONS	20
5.1 EXPLANATION OF THE RESULTS	23
6. CONCLUSION	23
7. ACKNOWLEDGMENTS	24
8. BIBLIOGRAPHY	25

TABLE OF FIGURES, CHARTS AND TABLES.....

TABLE 2,1.....5

FIGURE 3.1.....10

FIGURE 3.2.....11

FIGURE 4.1.....13

FIGURE 4.2.....14

FIGURE 4.3.....15

FIGURE 4.4.....16

FIGURE 4.5.....17

FIGURE 4.6.....18

FIGURE 4.7.....18

FIGURE 4.8.....19

CHART 5.1.....20

CHART 5.2.....20

CHART 5.3.....21

CHART 5.4.....21

CHART 5.5.....22

CHART 5.6.....22

1. Introduction

1.1 Abstract.

In this diploma work we developed a tool which can be used to evaluate the quality of machine translation systems, comparing the translations of MT with translations of native speaker. This application allows us to collect human judgment on translation output. For that to be possible we have implemented some tools such as 1) score the translation, 2) error classification, 3) fluency & adequacy of the translation, 4) mistranslated sentences. We have to say that the application has been developed to improve the machine translation system, although there are methods for automatic evaluation that reduce costs and runtime of evaluation process. Human evaluation is still necessary in MT research to improve these systems, since the results of automatic evaluation are not exact.

Keywords: machine translation, evaluation, applications, human evaluation of MT

1.2 Objectives.

The objective of this application is to evaluate the quality of the machine translation systems outputs to give some insights how the MT system can be improved. There are automatic methods such as WER, Meteor or BLUE which are used to evaluate the quality of MT systems automatically (using statistical methods). The main problem of the automatic evaluations is how to achieve a high level of precision in the measures obtained automatically with respect to the measures obtained on the same translation output manually by an annotator. Using automatic metrics we save cost and time, but the manual evaluation is still needed as it has a high level of correctness and gives detailed evaluation of the translation. However the manual evaluation is very time and cost consuming. To reduce this problem we have developed a graphical user interface (GUI) that allows annotators to get their own evaluation in a quick and simple way.

The development of an application to collect the human evaluation on machine translation output can be a complicated task. The annotator has to accomplish the following tasks: select translation score, classify errors, judge the fluency and adequacy of translations and select mistranslated sentences. Each annotator can have a different way of understanding the sentence. For this reason the

application has to have a design that can help to evaluate the translation that the annotator must do.

In this document we described how to use the application for human evaluation of MT and how to use the tools that allow collecting human judgments on translation output.

2. Machine Translation

2.1. Introduction to Machine Translation

The Machine Translation (MT) is a field in Computational Linguistics and uses the knowledge of other fields as well: informatics, linguistics, business, etc.

From 50's and beginning of the 60's of the XX century, there were some American engineers specialized in the artificial intelligence. They believed in the possibility of translating the texts automatically and that there would be a possibility that the machines would be able to do it. The MT started as a study that could be useful to reduce the translation costs of the companies and international organizations.

MT systems enable the translation of large bodies of text in a shorter time than a human is able to do. Projects such as “automatic translation of website” would be impossible without the help of machine translation systems. On the other hand, the MT also became a need of international organizations such as the European Community which has to generate many documents in different languages in a limited time. For this reason, the Community financed the project “Eurotrans” with the aim to develop a system able to translate the document automatically in all official languages of the European Union.

The MT is most successful when translating written documents in a controlled language. A document is written in a controlled language, if it has simple syntactic structures, and if it isn't ambiguous and has limited vocabulary.

2.1.1 Limitations

The limitations of a MT system alter the quality of the translation. If a MT system doesn't have an appropriate presentation of the source sentence meaning, it's most likely that the translation will have wrong meaning or will be illogical.

The comprehension of a sentence requires a complex knowledge of the source language and some elements to process the linguistic information. Obviously, the procedure of all this would cost a lot of effort and take a lot of time and probably the memory's resources of the system would collapse abruptly.

Nowadays, there is a high level of quality of the translations between Romance language (Spanish, Portuguese, Catalan, etc.). However, the results get worse when the languages are not similar, as it is in case of Spanish and English or German.

Another very influential point in the quality of the translation it's the degree of specialization in the translation systems. The quality of the translation can be improved if the translation system is specialized in a type of text and in a specific vocabulary. For example, a system specialized in the translation of weather reports will get high level of quality even to translate text between very different languages, but it will be useless to address, for example, sports or financial reports.

Translation is a hard task that requires a lot of knowledge and skills. In a translation it isn't enough to exchange one word for other, but you must also be able to recognize all the words in the context and the influence they have on each other. The human language has a specific morphology, syntax and semantics. So in the simplest text it can be a lot of ambiguities. It is also necessary to consider the matters of style, discourse and pragmatic.

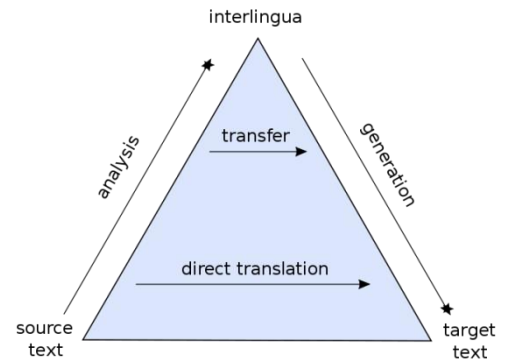
However, there are statistical methods that perform translations without taking the grammatical issues into consideration. Nowadays the trend is to integrate all kind of methodologies: explicit knowledge of language and statistics from a corpus.

2.1.2 Different kinds of Machines Translation

We can distinguish between two kinds of machine translation systems: rule-based and corpus-based.

- **Rule-based:** This type of machine translation systems is based on the principle of replacing words by their nearest equivalents. This kind of transformation of the source text is called pre-editing text.

Overall, in the first phase a text will be analyzed, usually the text is replaced with an internal symbolic representation. Depending of the abstraction of this representation, we can find different levels: from direct translation (making translations word by word) to interlingua (using a complete intermediate representation).



- **Corpus-based:** This machine translation system is based on the use of corpus that represents samples from real use.

Statistics: Nowadays the study in machine translation is centered on this systems because the results obtained are very promising. The costs and the time taken for its construction are lower than the cost of creation of translation engines with linguistic knowledge. Statistical machine translation tries to generate translations using statistical methods based on bilingual text corpora, such as the Canadian Hansard corpus, the English-French record of the Canadian parliament and EUROPARL[8], the record of the European Parliament. If such corpora are available, good results can be achieved translating similar texts, but such corpora are still rare for many language pairs. [7]

Example-based: It is often characterized by its use of a bilingual corpus as its main knowledge base, at run-time. It is essentially a translation by analogy and solves the translation problem by using similar solutions already resolved. [7]

Context-based: this machine translation system translates each word taking the words around it into account. The text is divided into units between four and eight words in length and these words are translated to target language, after the sentences parts without meaning are removed, then the word is moved for one position forward and the sentence is translated again, leaving only the sentence with meaning. This procedure

is repeated until the end of the text is reached. When all text has been analyzed, all results are joined in a sentence, obtaining unitary text.

	Advantages	Drawbacks
Rule-based	Good results in translation	Most complex, high investment in human capital
Corpus-based	Save costs and time	Poor result, if there aren't a large parallel corpus available

Table 2.1 Advantages and drawbacks of the different kind of machine translation.

2.2 Evaluation

The study in machine translation needs an appropriate judgment of obtained translations. Consistent and easy to use tools for the evaluation of the results are highly welcome for these tasks. It is needed to have some mechanism that can compare two different systems, or find out how any variation of MT system affects the quality of the translations.

The evaluation of a translation system presents a series of difficulties. First, it's a subjective process which is difficult to define. Frequently we find different approaches in the world of the machine translation.

The quality of a translation can be expressed in two principal attributes: the fidelity and the fluency of the text. While fluency is a monolingual evaluation, the fidelity is a bilingual evaluation, and therefore, it is a more costly process.

2.2.1 Measures “DARPA” of evaluation for translation

From 1992 until 1994, DARPA promoted a series of initiatives to define measures that evaluate the machine translation systems. As the result of those projects, three kinds of evaluation measures were introduced.

Adequacy:

The evaluators measure the correctness of the meaning in the output of the translation system in comparison to the meaning of the reference translation. For this reason, we show a translation of reference created by an expert to the evaluator (annotator), together with the translation created by a system being evaluated (google, bing, onelook, etc). The evaluator has a scale from 1 to N to evaluate the output. It is therefore a measure of fidelity.

Informativeness:

The evaluators answer multiple-choice questions about the translated text, as if it were a text analysis. It is another measure of fidelity.

Fluency:

The fluency measures the quality of a translation according to its degree of correctness in the target language, without taking into account the source sentence. The evaluators can see the proposed sentence and they have to evaluate it in a scale from 1 to N, in function of how it was accepted intuitively by a native speaker, besides that they have to consider the grammatical correctness and if the corpus of the translated text makes sense to the reader regarding the context.

2.2.2 Automatic Evaluation

In the automatic evaluation is not common to use the measures previously mentioned. In this case it is more frequent to use the objective measures that can be evaluated automatically. These measures serve as a reference to a possible translation for each of the sentence that we want to translate. This reference will be compared to the sentences proposed by the translation system. The most important measures are:

Word Error Rate (WER):

WER indicates the minimum percentage of words that have to be inserted, deleted or substituted in the translation to get the sentence of reference. It can be made automatically using the edit distance between two sentences. This measure can be calculated automatically and this can be a great advantage. Therefore, it is easy to get it and it is also reproducible (the result is always the same). The dependence with the sentence of reference is a big drawback in this measure. There is an almost unlimited number of correct translations for the same sentence and nevertheless, this measure considers that only one translation is correct.

Sentence Error Rate (SER):

This measure indicates the percentage of sentences whose translations do not match exactly with the expected reference sentence. This fact gives the same advantages and drawbacks that WER.

Some variants of WER were defined that can also be used automatically.

Position-Independent WER (PER):

The same that WER, but it considers any possible word order in the sentence of reference. Therefore, this measure doesn't take into consideration the capacity of a translation system to properly reorder the words in the output sentence. This method simply measures if the separate words have been generated, without taking into account their position in the sentence.

Multi reference WER (mWER)

The approach is identical to WER but takes several references into account for each sentence that was translated. For each sentence, the editing distance will be calculated with different references and the smallest result will be chosen. This approach has a drawback, because it needs a lot of human effort to introduce the references. Although, it can compensate the effort if a lot of evaluation will be made later.

Bilingual Evaluation Understudy (BLUE):

It's an automatic measure designed by IBM. This measure uses several references. The main trouble of mWER is the inability to translate all valid references. The measure "BLUE" tries to solve this problem combining the references that are available. In summary, we could say that BLUE measures the accuracy of the n-grams (unigrams, bigrams, trigrams and fourgrams) with respect to the set of reference translations.

The measure "BLUE" also includes a penalty for those translations whose length differs significantly from the length of the reference sentences.

2.2.3 Human Evaluation

Other kinds of measures have been developed where an intervention of one person to get the evaluation is necessary. Among the measure most often used, we could highlight the following:

Subjective Sentence Error Rate (SSER):

Each sentence is scored from 0 to N, regarding the quality of the translation. Here is an example of score:

- 0 – Without meaning.
- 1 – Some aspects of the content are transmitted.
- ...
- ...
- 5 – Understandable but with important syntactic errors.
- ...
- ...
- 9 – Ok, only slight errors of style.
- 10 – Perfect translation.

The biggest problem is the subjectivity, because two annotators can have different criterion to evaluate the same sentence. Another drawback is that different lengths of the sentences are not taken into account. The score of one sentence of 50 words has the same impact on the total score that a sentence of only 2 words.

Information Item Error Rate (IER):

This measure tries to solve the next issues: What do we have to do if there is a long sentence and there are parts of the sentence with a correct translation and others parts of the sentence with an incorrect translation?. To solve the problem, the concept “information items” is introduced. The sentences are divided into segments of words called “information items”. Each item of the input sentence is qualified as “ok”, “fail”, “syntactic” or “others”, it depends on the translation. The measure IER can be calculated as the percentage of the items mistranslated (no qualified “ok”).

Information Item Semantic Error Rate (ISER):

ISER is a modification of IER, where an item is considered correct, if the desired information is transmitted, without taking possible syntactic errors into account.

3. Development System

In the continuation of the report the system for human evaluation of MT outputs will be described. First we will describe the development environment, and the application we have designed.

3.1 Development Tools

3.1.1 Java SE

Java programming language is a high-level language with concurrence. It is class-based and object-oriented. The applications created with Java don't depend on the hardware and allow “to program once and to run in different sites”.

This feature makes Java appropriate programming language to corporate and internet applications, where we can find different hardware platforms: Windows, Linux, Unix, Mac, etc. When we compile the program in Java, an independent code is generated in the same place that the code was created. This code is known as bytecode and this bytecode is interpreted in the computer on which it is run.

For that being possible, it is necessary that the computer can interpret the code on which the bytecode is run. For this reason, the computer has to have what we know as a virtual machine of Java (JVM). The virtual machine of Java is not installed by default in the computer, we have to install it. One of the ways to do that is given in the website of Oracle.

Sun Microsystem divided Java in three big branches, each of them with their set of APIs and their own development tools: big computers, desktops and microcomputers or dispositive of limited memory.

Java SE is the edition for desktops and the platform that we will use to develop our application.

3.1.2 Development Environment

The Java technology is closely related to the world of “Open Source” and this is one of the advantages of Java. For this reason it is easier to find a lot of free IDEs. One of them is NetBeans. This IDE is offered by Sun and it's the environment that we will use in the creation of our application.

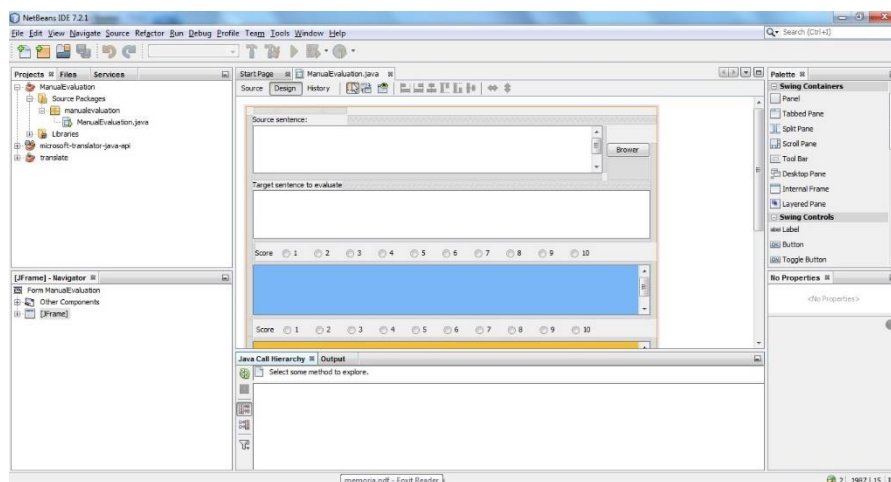


Figure 3.1 NetBeans Interface

4 Tools for Human Evaluation of MT

In our tool for Human Evaluation of MT we have used different methodologies to measure the quality of the machine translation, which we will explain later. We will start with a short explanation about how to use our application for human evaluation of MT.

4.1 How to use the application for human evaluation of MT

4.1.1 How to open the files

The annotator first selects a file of reference (using the button “Open”), where the source sentences (sentences that we have to evaluate) will be stored with the possible translations of reference (sentences that have been translated by MT system).

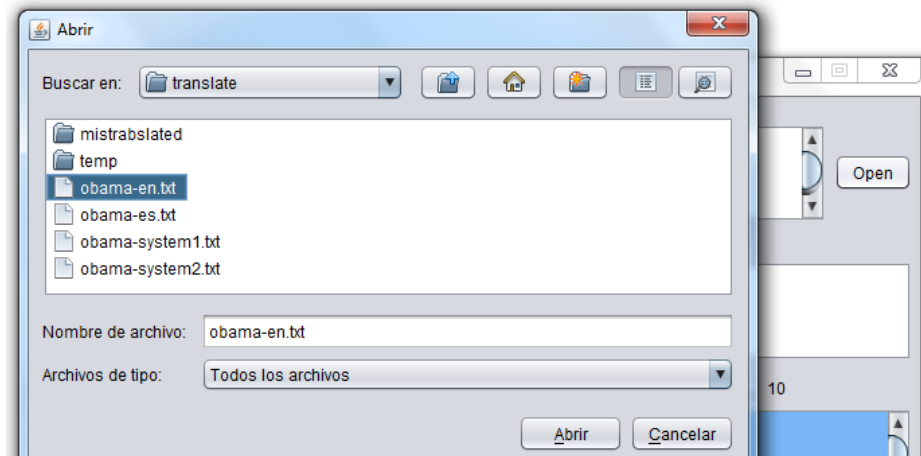


Figure 3.2 Selection of the files.

We only have to select for example the file “obama-en.txt” and the other files, named appropriately will open automatically (“obama-es.txt, obam-system1.txt, Obama-system2.txt”). Next we will explain the content of the files that we see in the screenshot:

- obama-en.txt: this file has all sentences in the source language (in the example it was English) that the annotator will have to compare with the sentences of MT.
- obama-es.txt: this file contains all translations of “Obama-en.txt”, which were made by a native speaker. This file helps to evaluate the translations of MT easier.
- obama-system1-txt: this file has the translations, obtained with “System 1” MT.
- obama-system2-txt: this file has all translations of the “System 2” MT.

These files will be stored in the directory: \NetBeansProjects\ManualEvaluation.

4.1.2 Parts of the application

The application can be separated in five parts that we will describe in some details below:

Frames:

In this part we can distinguish between four different frames where we can see different translations of MT and the source sentence together with the translations of native speaker. In the first frame there will be the source sentence, in the second frame we will see the translation by a native speaker, in the blue frame there will be the translation of “System 1” MT, while in the orange frame there will be the translation of “System 2” MT.

Translation score:

In this part we use the method “Subjective Sentence Error Rate” to evaluate the translation of MT. The method SSER was described in the section “Human Evaluation”.

We have used a score from 1 to 10, where the worst translation will be evaluated with a 1 and the best translation will be evaluated with a 10.

Error Classification:

In this part we have developed a tool to analyze the errors that the translation can have. We have created a section where the annotator has to classify the errors present in the given translation. The errors that we have considered more important for a translation are:

- **incorrect word form(s) ,**
- **incorrect word order ,**
- **content word(s) wrong in meaning , and**
- **missing content word(s) .**

Fluency & Adequacy:

With this tool we can measure the fluency and the adequacy of a sentence. The annotator has to evaluate the fluency and adequacy of the translation taking into account the translation of the native speaker.

This measure was previously mentioned in the section of “Measures DARPA of evaluation for translation”.

Mistranslated sentences:

In this part of the application the annotator can select the part of the mistranslated sentence produced by MT system. Once this is done, the annotator has to write the correct translation in the frame.

We can see different parts of the application in the next figure.

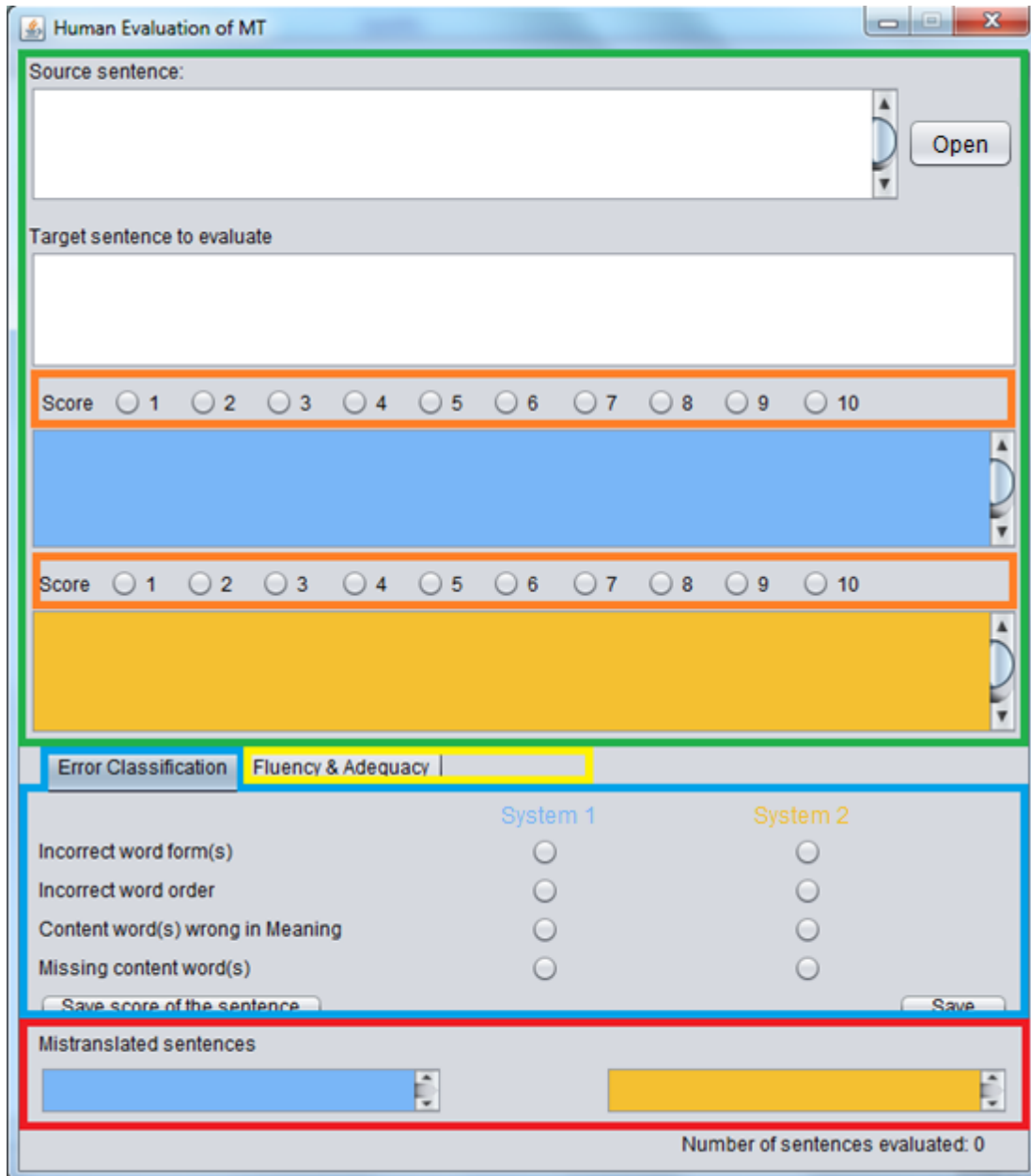


Figure 4.1 Parts of the application “Human Evaluation of MT”

4.1.3 How to save the evaluation of the annotator.

Once the annotator has opened the file “Obama-en.txt” and the sentences are ready to evaluate, the annotator should start the evaluation procedure. When the annotator has done the evaluation of the first sentence, he must push the button “Save score of the sentence” and the application will automatically proceed to the next translation (the score will be refreshed to avoid confusion about the sentence evaluation). When the annotator has finished evaluating all the sentences, all frames will be cleaned and the annotator will have to push the button “save” (to save all data). With this button the annotator will generate two different documents (.txt) with all results and statistics of human evaluation of each MT (System 1 and System 2) [Figure 4.3].

The temporary files will be saved in the directory “...\NetBeansProjects\ManualEvaluation\translate\temp”, while the files “mistranslated” will be saved in the directory “...\NetBeansProjects\ManualEvaluation\translate\mistrabslated”.

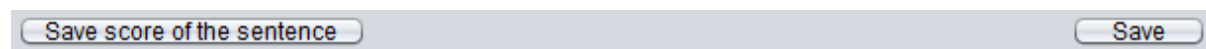


Figure 4.2 Buttons to save the evaluation of current sentence

4.1.4 Results (explication)

As we have previously mentioned, when the annotator saves all evaluations of the translations, the application generates a document with the results of human evaluation of each MT. Below we will explain the results of this example:

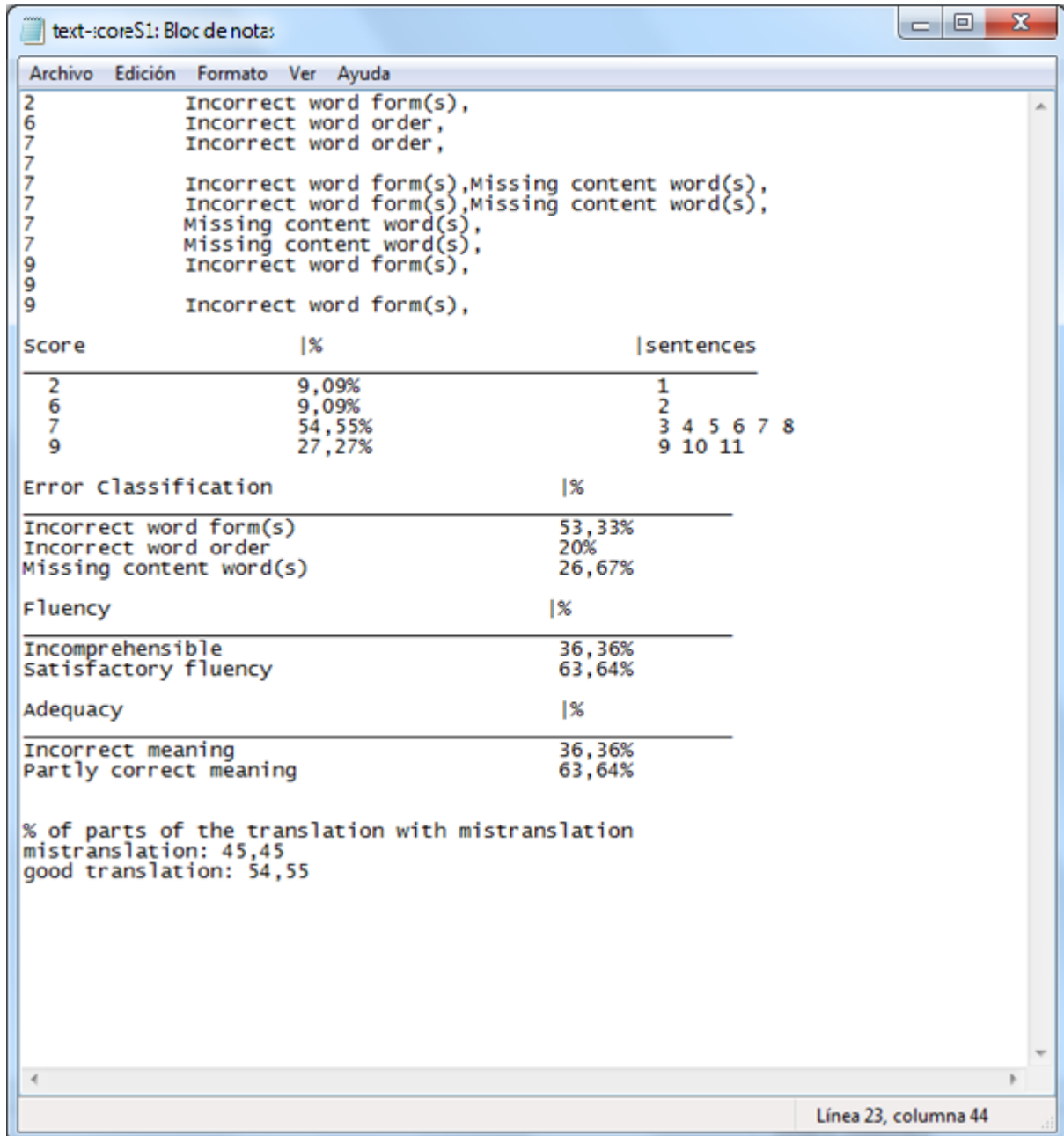
We can distinguish between five big sections in the document that has been generated:

The first section of the document is separated in two columns. In the first column the score of the sentence are displayed and in the second column we can see different “errors of classification” that the sentence can have. The score and the errors classifications of the first sentence will be located in the first line of the document, the score and errors classifications of the second sentence will be located in the second line and so forth.

The second big section will be the score of each sentence and the percentage of each score.

The third and fourth sections contain different evaluations of fluency and adequacy and the percentage of each evaluation.

Finally, in the fifth section we have the percentage of the sentences with a correct translation and percentage of the sentences with mistranslation.



```

2      Incorrect word form(s),
6      Incorrect word order,
7      Incorrect word order,
7
7      Incorrect word form(s),Missing content word(s),
7      Incorrect word form(s),Missing content word(s),
7      Missing content word(s),
7      Missing content word(s),
9      Incorrect word form(s),
9
9      Incorrect word form(s),

```

Score	%	sentences
2	9,09%	1
6	9,09%	2
7	54,55%	3 4 5 6 7 8
9	27,27%	9 10 11

Error Classification	%
Incorrect word form(s)	53,33%
Incorrect word order	20%
Missing content word(s)	26,67%

Fluency	%
Incomprehensible	36,36%
Satisfactory fluency	63,64%

Adequacy	%
Incorrect meaning	36,36%
Partly correct meaning	63,64%

% of parts of the translation with mistranslation
mistranslation: 45,45
good translation: 54,55

Figure 4.3 Final document with all results of the human evaluation of MT

4.1.4.1 Tools to evaluate the translation

In our application we have used four tools to evaluate the translations of MT system, “score of the translation”, “error classification”, “fluency & adequate” and “mistranslated sentences”. Below we will explain each one of them:

Score of the translation: this tool was developed to evaluate the quality of the translation. We have used a scale from 1 to 10, 1 being the worst score and 10 the best score. The annotator has to evaluate the translation, taking into consideration the grammar, morphology, syntax and semantics. Below we show a screenshot of this tool.

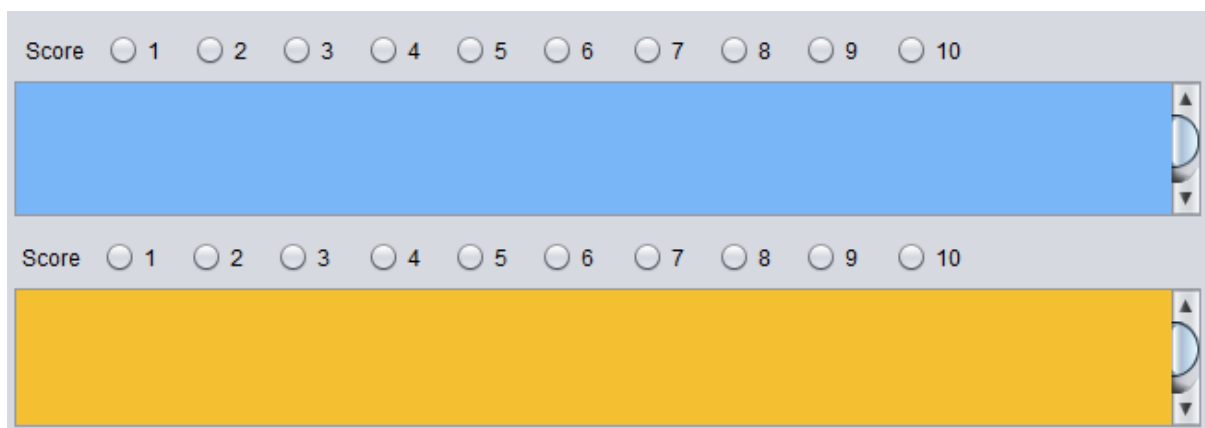


Figure 4.4 Score of the translation

Error Classification: with this tool we measure different kinds of errors that the translation can have. This must be done by the annotator. Below we will give a short explanation about the four errors of classification that we have considered the most important for a translation.

Incorrect word form(s): with this error we will know if the translation system selected correct word form of a given word. An example of an incorrect word form is:

This is my **final** vacation plan: I'm going to San Francisco for a week.

Este es mi **último** plan de vacaciones: Me voy a San Francisco durante una semana.

Este es mi plan de vacaciones **definitivo**: me voy a San Francisco durante una semana.

The MT have translated **final** as **the last** when the MT should translate the word as **definitive**.

Incorrect word order: this error helps us to know if the translation has a correct word order. This kind of error is very important because an incorrect word order can change the meaning of the sentence.

I was shopping in Leipzig -- sentence to translate

Yo estaba de compras en Leipzig. -- correct

Yo estaba **en Leipzig** de compras. -- incorrect

in English the translation would be:

I was **in Leipzig** shopping -- incorrect

Content word(s) wrong in meaning: with this error we can get the information if the translated words have different meaning that the words should have.

Missing content word(s): often, the machine translations lose words in their translations, leaving the translation without meaning. With this error we will know if the translations lost words in their translations.

Error Classification	Fluency & Adequacy
	<div style="display: flex; justify-content: space-around;"> Google Bing </div>
Incorrect word form(s)	<input type="radio"/> <input type="radio"/>
Incorrect word order	<input type="radio"/> <input type="radio"/>
Content word(s) wrong in Meaning	<input type="radio"/> <input type="radio"/>
Missing content word(s)	<input type="radio"/> <input type="radio"/>
<input type="button" value="next sentence"/>	<input type="button" value="Save score of the sentence"/> <input type="button" value="Save"/>

Figure 4.5 Error Classification

Fluency & adequacy: with this tool we can measure the fluency and adequacy of the translation. This two scoring models have a four-way scoring. The four-way scoring of fluency is: “incomprehensible”, “bad fluency”, “satisfactory fluency” and “excellent fluency”, while in the adequacy the four-way scoring is: “nonsense”, “incorrect meaning”, “partly correct meaning” and “correct meaning”.

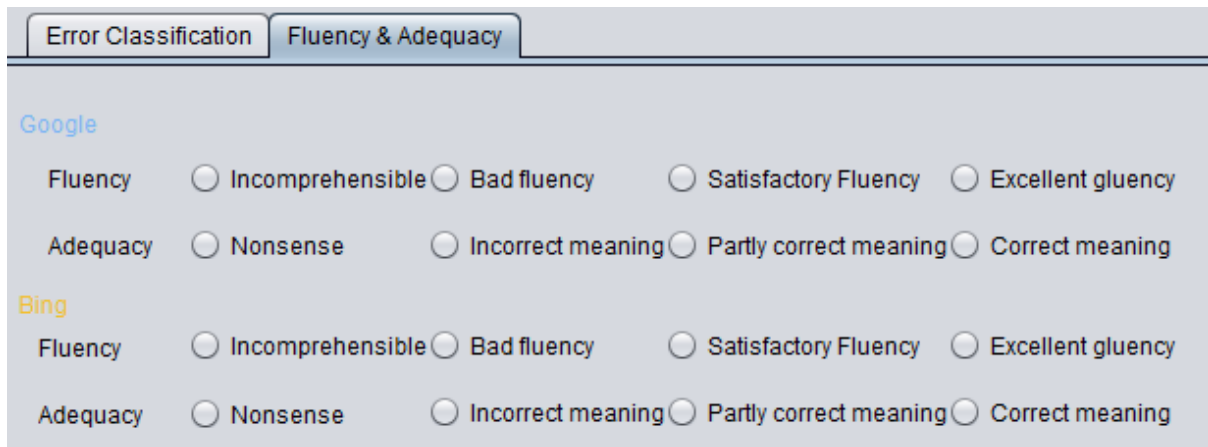


Figure 4.6 Fluency & adequacy

Mistranslated sentences: often, when the MT system has to translate a long sentence, there are parts of the translation that are correct and other parts of the translation that are mistranslated. With this tool we try to solve this problem. The annotator must select the part of the translation that has a mistranslated part and write down the translation that he/she considers correct (there will be one frame for each MT). The application will generate two files “MistranslatedS1” and “MistranslatedS2” [Figure 4.8]. This files will be saved in the directory ...\\NetBeansProjects\\ManualEvaluation\\translate\\mistrabslated. In these files we can see the mistranslated sentences selected by the annotator and the sentence that the annotator considered correct. When the annotator finishes the evaluation of all the sentences, the percentage with the correct translations and mistranslations will be put in the file where we have saved all the scores [Figure 4.3].

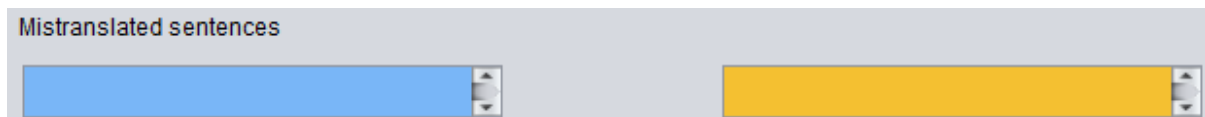


Figure 4.7 Frames where the annotator must put good translation

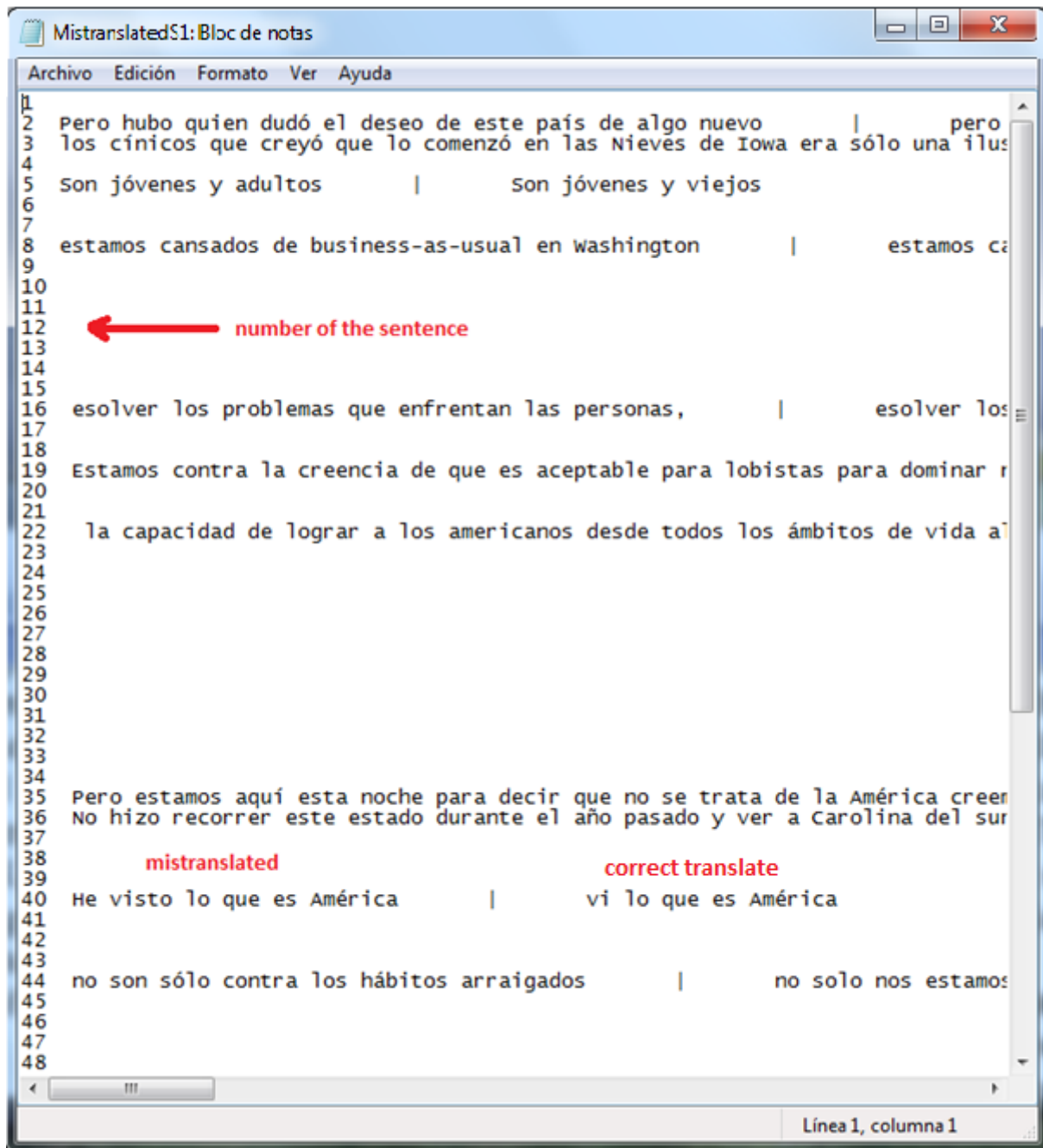


Figure 4.8 File with the mistranslation and correct translation of each sentence

Note: if the line of the sentence is in white it's because the translation was correct.

5. The results from the evaluation of the translations

This section contains the summary of the human evaluation.

SCORE OF TRANSLATION

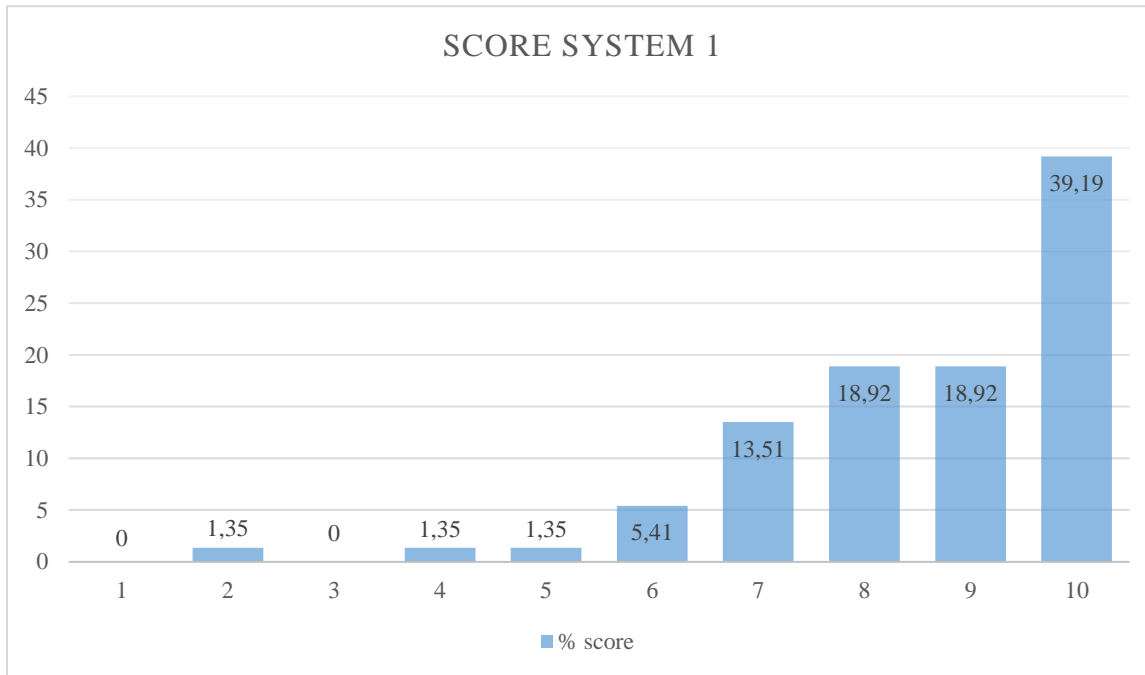


Chart 5.1 Score of System 1

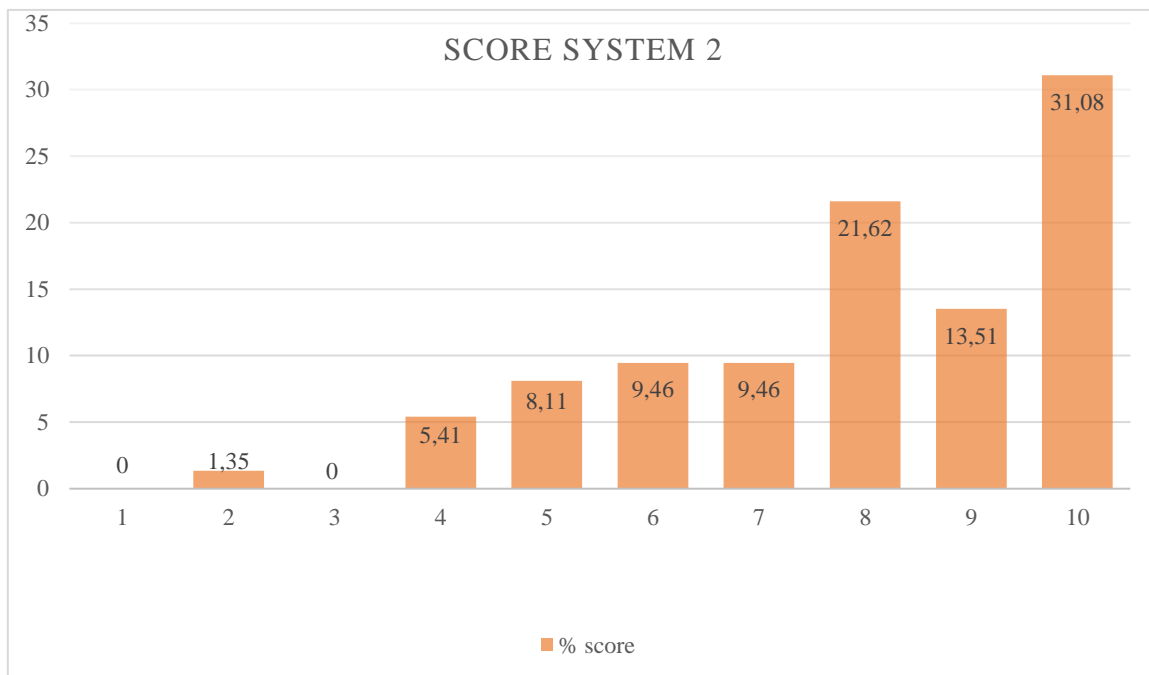


Chart 5.2 Score of System 2

ERROR CLASSIFICATION

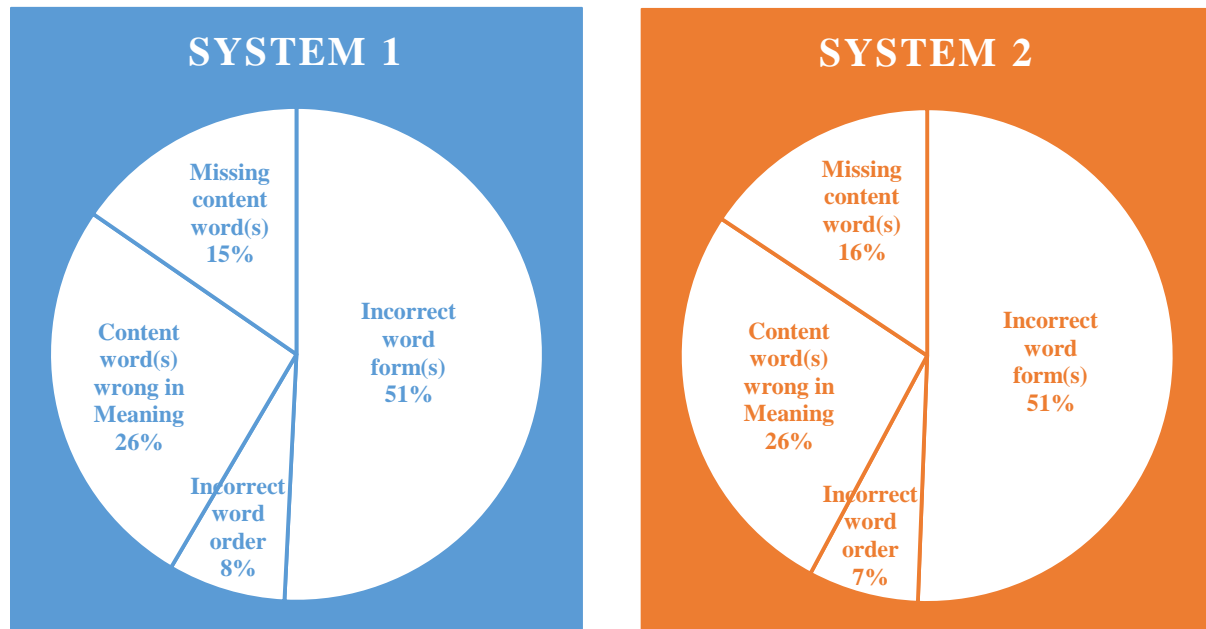


Chart 5.3 Error Classification of System 1 and System 2

FLUENCY

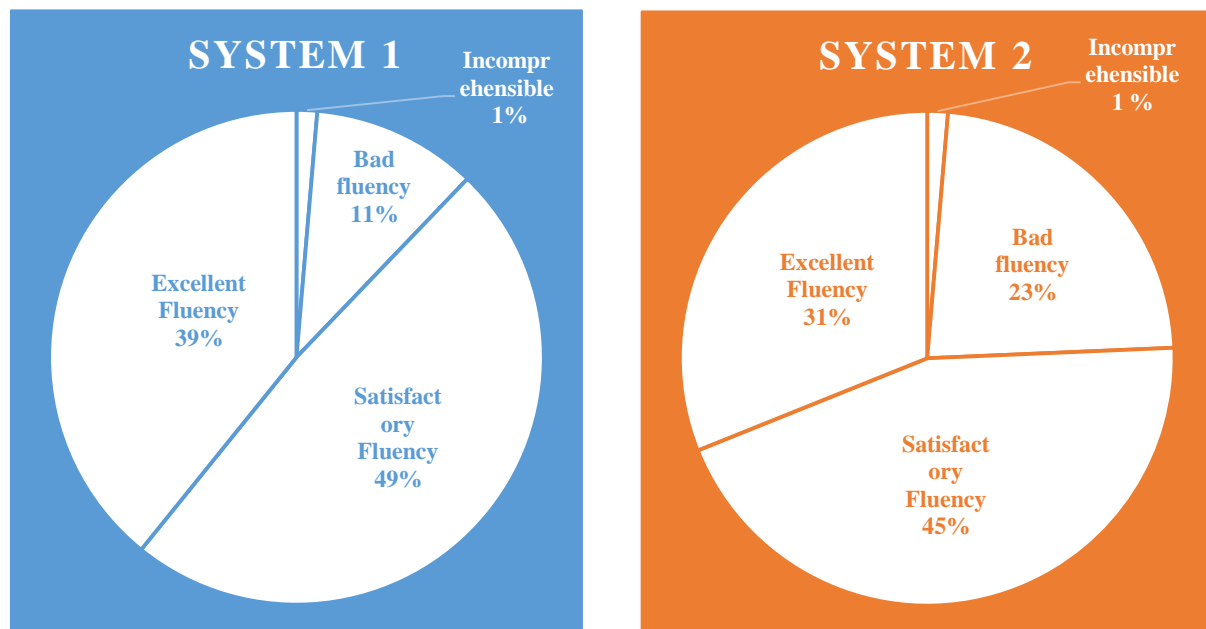


Chart 5.4 Fluency of System 1 and System 2

ADEQUACY

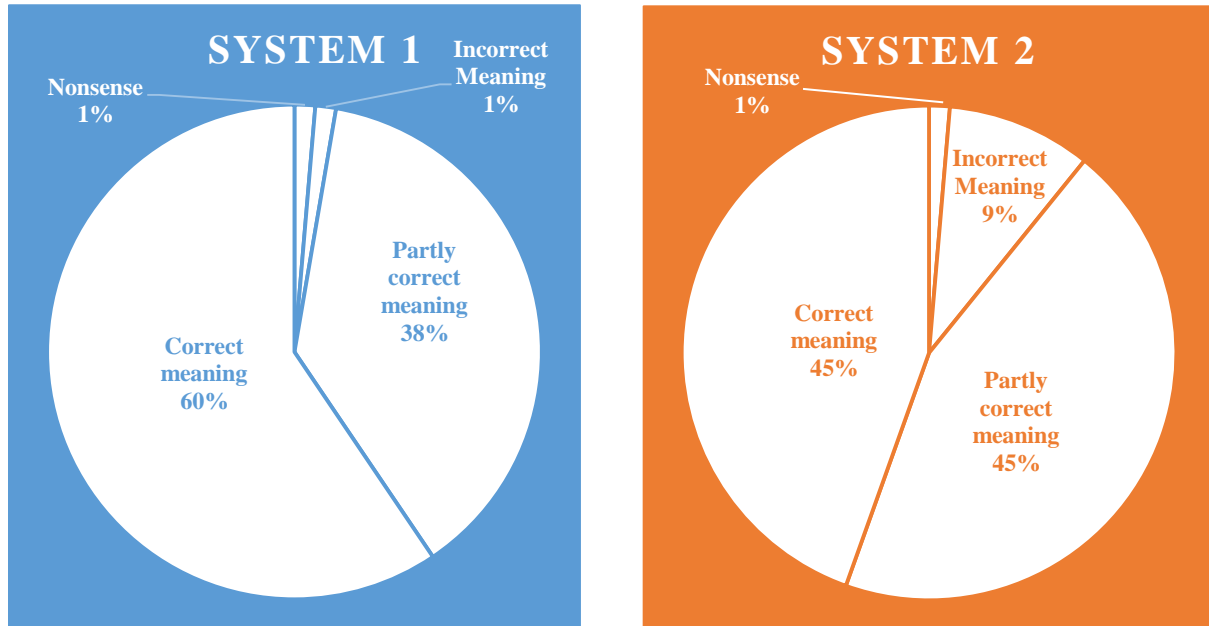


Chart 5.5 Adequacy of System 1 and System 2

MISTRANSLATED SENTENCES

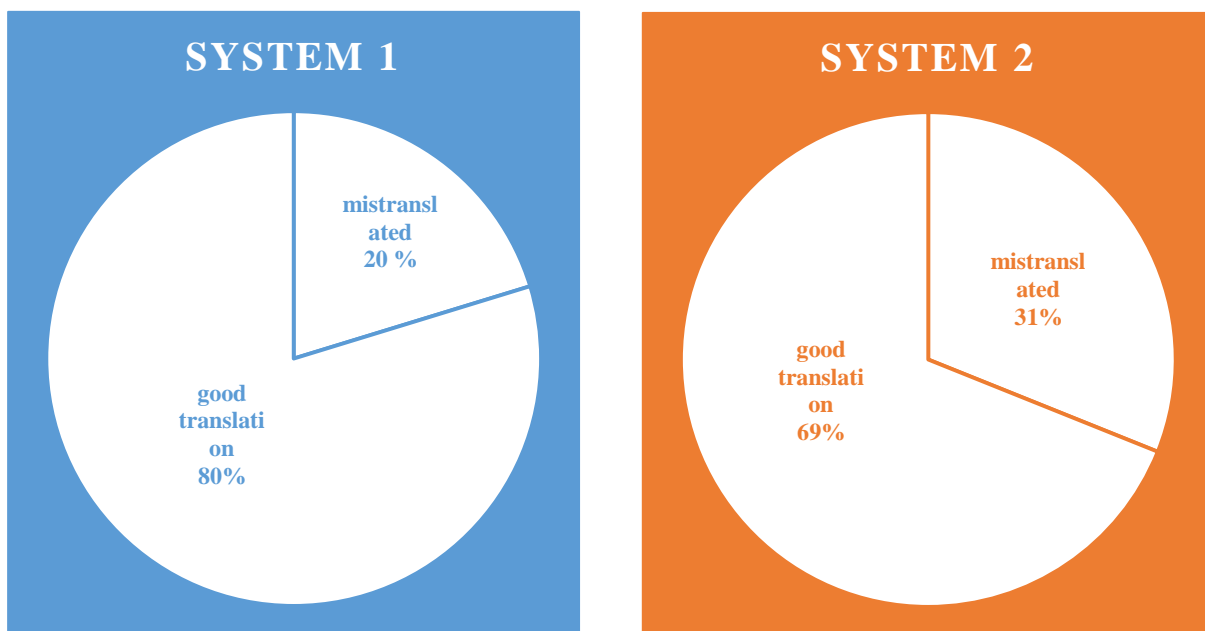


Chart 5.6 Mistranslated sentences of System 1 and System 2

5.1 Explanation of the results

If we look at the chart 5.1 and the chart 5.2 we can say that in general both System 1 and System 2 produce good translations of the sentences, although the translation of System 1 are in average better that the translations of System 2. As we can see comparing chart 5.1 and 5.2, the percentages with good translations (scores 7, 8, 9, 10) in chart 5.1 are higher that the percentages of good translations in the chart 5.2. However, in the chart 5.2 the percentages with a score 4, 5 and 6 are higher that the percentages from “System 1” MT. Finally we can say that both “System 1” MT and “System 2” MT make a good translation but the “System 1” MT has a machine translation system that translate the sentence better than “System 2” MT.

In the chart “error of classification” [Chart 5.3], we can see both System 1 and System 2 have the same errors and the most common error is “Incorrect word form(s)” where more than 50% of the sentences have this kind of error.

As we can see in the chart 5.4 the fluency for both machine translation systems are quite acceptable, as the 75% of the sentences have an “excellent fluency” or a “satisfactory fluency”. In the other hand, we can see in the chart 5.5 that the adequacy of the System 1 MT is almost perfect, with the 59,46% of the sentences with a correct meaning and the 37,84% of the sentences with a partly correct meaning, while in the case of the System 2 MT the percentage of "correct meaning" and "partly correct meaning" is 45% for both.

Finally, we can see that the System 1 MT has less mistranslations that System 2 with 80% of sentences with a good translation while the System 2 MT has 69% of the sentence with a good translation, which is 11% less than the sentences by System 1.

At the end we must point out, that this evaluation was just an example. To get the general conclusion about the preferences of one system over the other, the evaluation set must be large.

6. Conclusion

In this paper, we have described the application “Human evaluation of MT” where we have developed tools such as “score of the sentence”, “error classification”, “fluency & adequacy” and “mistranslated sentence”. The development of these tools couldn't have been possible without the help of reports such as “Appraise: an Open-Score Toolkit for Manual Evaluation of MT Output”[1,2] or “(Meta-) Evaluation of Machine Translation”[4]. These reports

talk about machines translation, evaluation of MT (automatic and human) and the different application that have been created to develop these technologies. These reports introduced me in the world of the MT. Mainly I have developed my own application on the basis of the application “Appraise”[1,2] by Christian Federmann, he allowed me to use "Appraise" to know how the human evaluation of MT works, when I started to use his program I started to understand how to work the toolkit for manual evaluation on MT should work and then I started to develop my own application on the basis of "Appraise".

I have to say that this application is only a first version, and we could add other task such as Information Item Error Rate or Information Item Semantic Error Rate to improve the quality of the evaluation (these measures were previously mentioned in the section Human Evaluation). We also can add a tool where the annotator can choose different MT system to evaluate. From my own experience as an annotator, I think that it is better to evaluate MT systems separately. This will avoid confusion when we will evaluate the sentence.

7. Acknowledgments

I have to give thanks to my coordinator Mirjam Sepesy Maučec for her help to get information about MT and the steps that I have had to do to develop my diploma work, I also have to give thanks to Christian Federmann for his help to use his application (Appraise) and last but not least, I have to give thanks to my family and my friends for their support.

8. Bibliography

[1]. Christian Federmann.: Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output. The Prague Bulletin of Mathematical Linguistics No. 98, 2012, pp. 25–35. *Language Technology Lab, German Research Center for Artificial Intelligence, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany.*

[2]. Federmann, C.: Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valetta, Malta (May 2010), http://www.lrec-conf.org/proceedings/lrec2010/pdf/197_Paper.pdf. *Language Technology Lab, German Research Center for Artificial Intelligence, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany.*

[3]. Christian Federmann: How can we measure machine translation quality? Proceedings of the Tralogy Conference, Paris, France, L'Institut de l'Information Scientifique et Technique, 12/2011.

<http://odel.irevues.inist.fr/tralogy/index.php?id=76&format=print>

[4]. Chris Callison-Burch, Cameron Fordyce, Philipp Hoehn, Chistof Monz, Josh Schroeder.: (Meta-) Evaluation of Machine Translation. Proceedings of the Second Workshop on Statistical Machine Translation, pages 136–158, Prague, June 2007

[5]. Wikipedia, Métodos de evaluación para la traducción automática. http://es.wikipedia.org/wiki/Métodos_de_evaluación_para_la_traducción_automática

[6]. Wikipedia Traducción automática. http://es.wikipedia.org/wiki/Traducción_automática

[7]. Wikipedia, Machine translation. http://en.wikipedia.org/wiki/Machine_translation

[8]. www.statmt.org/europarl

[9]. Jesús Tomás, Josep Ángel Mas, Francisco Casacuberta,: A Quantitative Method for Machine Translation Evaluation. Proceedings of the EACL Workshop Evaluation Initiatives in Natural Language Processing, Budapest, Hungary, pp. 27-34, 2003

[10] Java Web Start Guide.

<http://docs.oracle.com/javase/6/docs/technotes/guides/javaws/developersguide/contents.html>