# AUTOMATIC MACHINE TRANSLATION USING COMPARABLE CORPORA

Catarina Moreira and Pedro Gonçalves

Instituto Superior Técnico, TagusPark – Universidade Técnica de Lisboa
Av. Prof. Dr. Aníbal Silva 2744-016 Porto Salvo
catamoreira@msn.com, prsg2000@gmail.com

**Abstract:** With this work, we propose ourselves to search for similarity metrics and techniques that permit the extraction of a bilingual lexicon of the Portuguese and English languages. We also propose an algorithm, based on the researched techniques, in order to improve the lexicon extraction. The extraction of bilingual lexicon is based mainly on the similarity of the words. Our algorithm achieved an F-measure of 91, 65%.

**Keywords:** comparable corpora, extraction, lexicon, Jaro, Jaro-Winkler, Longest Common Substring, Minimum Edit Distance, Soundex, similarity, frequency, identical words.

## 1. Introduction

For some time, people have established progressively closer connections between the various countries around the world. However, one of the biggest barriers preventing the flow of information and the share of knowledge from those connections is the language.

Machine translation has been the key technique to bridge the language barrier.

For this reason, the main objective of this work is the extraction of bilingual lexicons, mainly based on the similarity of the words, using comparable corpora.

Our work is divided into two phases: (1) research, where we will experiment and analyze what are the best metrics that are more adequate for the bilingual lexicon extraction and (2) an algorithm proposal, based on the conclusions drawn from the research carried out, in order to obtain better results.

For the experiments conducted, in a first step, we used one of the seventy eight test file corpus, analyzed the results and took some conclusions.

At the end of this work, we make a general analysis of all the techniques against our algorithm, but for four test corpus files.

As tools, we used the Qizx studio which is a fast search engine and repository for XML. This allowed the rapid collection and observation of results. We also used libraries that were integrated into Qizx studio, which had some of the researched techniques already implemented.

## 2. State of Art [8]

There aren't many techniques that can extract bilingual lexicons from comparable corpora, compared to the techniques that are used for the extraction of lexicons in parallel corpora.

Currently, the strategies that are being explored are based on the context similarity. A word w2 in a given language is a possible translation of a word w1 of another language, if the expressions of context with which co-occurs w2 are translations of the terms of the context that co-occur with w1.

The main objective of this strategy is to find words in a target language that have distributions that are similar to the words in the source language. One way to start this strategy is through bilingual lists of expressions that are used to construct context vectors of all words of both languages. This list is usually provided through bilingual dictionaries. However, there are many approaches that experienced to put other sets of words in that initial list.
Anyway, in all cases the list contains the "seed words" necessary for the construction of context vectors.
There are also other approaches that do not use this initial list. However, the results of these approaches were not good enough to be accepted. One of the main differences that exist in these strategies is in the coefficients used to measure the similarity of the vectors. Another difference is the

way they define the contexts of words. Most related work, define the context as a window of words of size N.

We have not explored this approach of context similarity, for being too complex. The main objective of this work, as already mentioned, is to propose an algorithm, using the most adequate metrics to detect the similarity between two words.

## 3. Researched Techniques

This section describes techniques for finding possible translations of words between two corpus of different languages.

The techniques considered are:

- **Identical Words** – two languages may share a certain number of identical words. For example: *chocolate* and *email*.
- **Similar Spelling** – due to the evolution of natural languages, some words may undergo certain changes from its root word. For example: *organização* and *organization* (only their suffixes are different).
- **Similar Sound** – in two different languages, there may be words that have different spelling. However, by comparing them with their phonetic, they are extremely similar. For example: *líder* and *leader*.
- **Word Frequency** – when dealing with comparable corpora, if one word occurs many times in a corpus, then its translation should also occur several times in the other corpus.

Thereafter, we intend to check in detail, in what sense is that these techniques contribute to the construction of bilingual lexicons between the Portuguese and the English language.

### 3.1. Identical Words

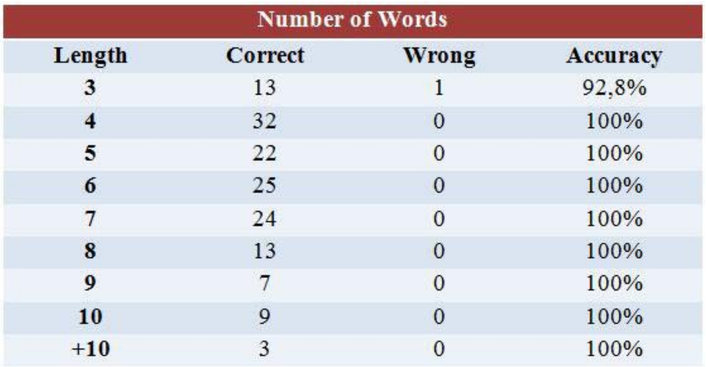
To handle identical words, we analyzed the Portuguese words contained in our corpus and looked for the exact same ones in the English corpus.

However, this has not yielded great results. Most of the words found, consisted in personal names (which are always the same for all languages) or in company names and not many common names.

These results were already expected.

Given that Portuguese is a Latin language and English is Germanic, then we believe that it is normal not to find lots of words spelled exactly the same way in these languages. These languages have different origins.

The table below shows the results of this test:



| Number of Words | | | |
|---|---|---|---|
| **Length** | **Correct** | **Wrong** | **Accuracy** |
| **3** | 13 | 1 | 92,8% |
| **4** | 32 | 0 | 100% |
| **5** | 22 | 0 | 100% |
| **6** | 25 | 0 | 100% |
| **7** | 24 | 0 | 100% |
| **8** | 13 | 0 | 100% |
| **9** | 7 | 0 | 100% |
| **10** | 9 | 0 | 100% |
| **+10** | 3 | 0 | 100% |

*Table 1: test results using identical words.*

Although the results are mostly names, we also found words common to both languages. And we only found one word mistranslated.

This leads us to conclude that one word in Portuguese written in the exactly same way as a word in English, suggests that these words are potential translations of each other.

### 3.2. Word Frequency

When analyzing two comparable corpora, one can see that the words which occur more frequently in a corpus will certainly refer to the same concepts in the other one.

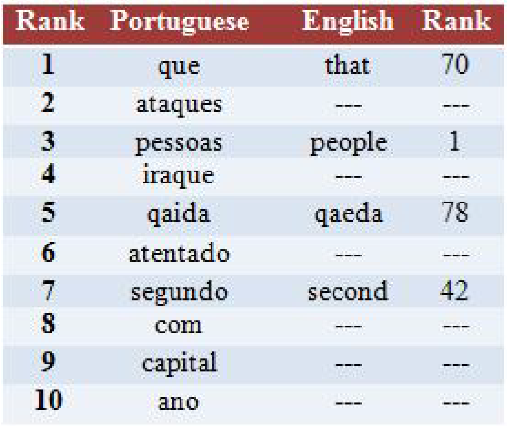
To evaluate this, aligning the nth Portuguese word with the nth English word, would not be a viable option, because it will lead to incorrect results. For example, in the English language there are lots of auxiliary words and verbs (such as *do*, *does*). In the Portuguese language there aren’t that many. So when comparing the nth Portuguese word with the nth English word, there will be mistranslations.

Because of that, we developed a formula to calculate the frequency of a word. The frequency is all occurrences of a certain word that occur in a document.

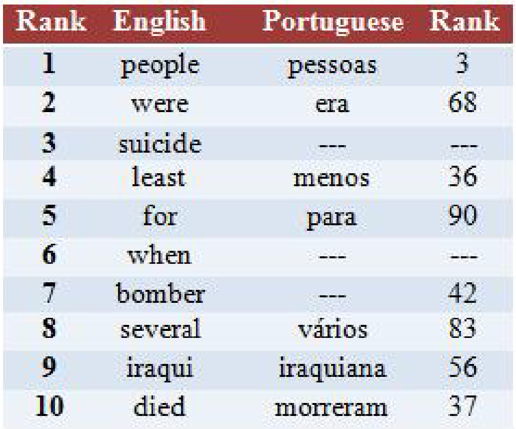
$$Freq(w) = \frac{count(w)}{corpusSize}$$

We thought in this measure, for a simple reason: a word that occurs several times in a small corpus is certainly very relevant. In the other hand, if a word occurs some times in a big corpus, then it isn't important at all.

The table below shows the results of this test:



| Rank | Portuguese | English | Rank |
|---|---|---|---|
| 1 | que | that | 70 |
| 2 | ataques | --- | --- |
| 3 | pessoas | people | 1 |
| 4 | iraque | --- | --- |
| 5 | qaida | qaeda | 78 |
| 6 | atentado | --- | --- |
| 7 | segundo | second | 42 |
| 8 | com | --- | --- |
| 9 | capital | --- | --- |
| 10 | ano | --- | --- |



| Rank | English | Portuguese | Rank |
|---|---|---|---|
| 1 | people | pessoas | 3 |
| 2 | were | era | 68 |
| 3 | suicide | --- | --- |
| 4 | least | menos | 36 |
| 5 | for | para | 90 |
| 6 | when | --- | --- |
| 7 | bomber | --- | 42 |
| 8 | several | vários | 83 |
| 9 | iraqui | iraquiana | 56 |
| 10 | died | morreram | 37 |

*Table 2 and 3:the frequency ranks of the most frequent Portuguese and English words and their translations.*

The "---" in the table represents that no translation was found in the corpus. Since we are dealing with comparable corpora, there was a big risk of not finding translations of many words, because they were not used in the corpus.

### 3.3. Word Similarity

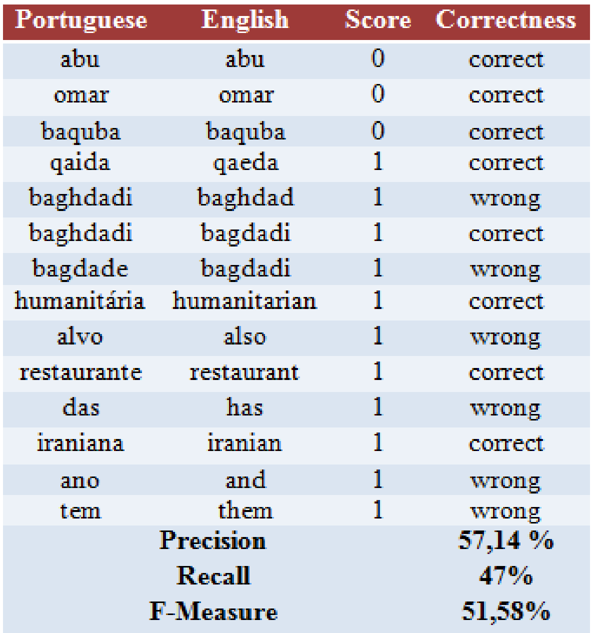
Due to the influence of people who passed through Portugal, particularly the British and the French, these influences led to some changes in the Portuguese language.

That is, some words became more similar to the words in English. For example: *police* and *polícia*, *conclusion* and *conclusão*, and so on.

These differences mainly differ in their suffixes.

In this section, we will describe some metrics that we researched and which are used to compare word similarity:

**3.3.1. Minimum Edit Distance [1]**: is a metric to measure the amount of different characters between two sequences. The edit distance between two strings is given by the minimum number of operations needed to transform a string into another. The results obtained, using this metric, are shown in the table below.



| Portuguese | English | Score | Correctness |
|---|---|---|---|
| abu | abu | 0 | correct |
| omar | omar | 0 | correct |
| baquba | baquba | 0 | correct |
| qaida | qaeda | 1 | correct |
| baghdadi | baghdad | 1 | wrong |
| baghdadi | bagdadi | 1 | correct |
| bagdade | bagdadi | 1 | wrong |
| humanitária | humanitarian | 1 | correct |
| alvo | also | 1 | wrong |
| restaurante | restaurant | 1 | correct |
| das | has | 1 | wrong |
| iraniana | iranian | 1 | correct |
| ano | and | 1 | wrong |
| tem | them | 1 | wrong |
| | **Precision** | | **57,14 %** |
| | **Recall** | | **47%** |
| | **F-Measure** | | **51,58%** |

*Table 4: test results using minimum edit distance metric.*

These results are for a threshold < 2. We also tested with other thresholds bigger than 2.

What we observed was that the higher the threshold used, the greater the number of pairs of words found. For bigger thresholds, the recall percentage is high; in contrast the accuracy returns too poor results. The minimum edit distance is very useful when leading with words misspelled. For cases like: *organization* and *organisation*, it would be a perfect metric. However, in the context of machine translation, it didn't return good results, so this isn't a very relevant metric in this context. This is the conclusion we take from these results.

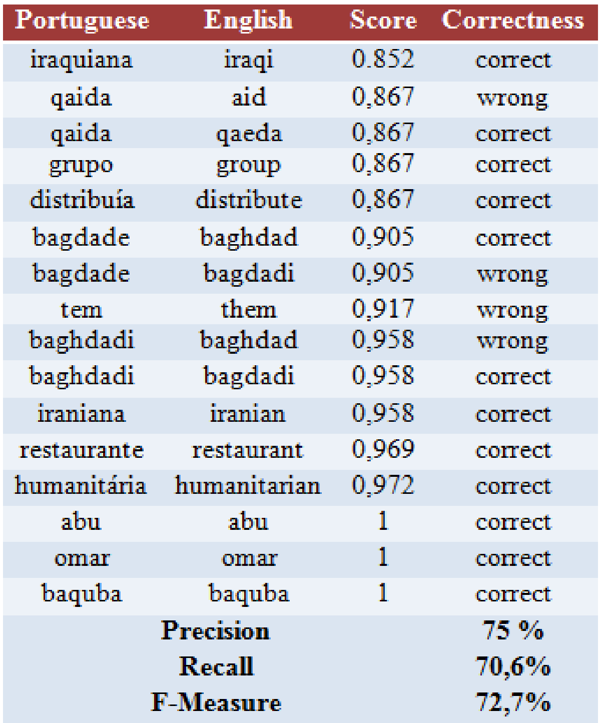
**3.3.2. Jaro Distance**: is a metric that takes into account the spelling deviations that typically occur between two words.

Briefly, for two strings *s* and *t*, let *s'* be the characters in *s* that are "common with" *t*, and let *t'* be the characters in *t* that are "common with" *s*. Roughly speaking, a character *a* in *s* is "in common" with *t* if the same character *a* appears in about the place in *t*.

Let *Ts,t* measure the number of transpositions of characters in *s'* relative to *t'*. The Jaro similarity metric for *s* and *t* is: [2]

$$Jaro(s,t) = \frac{1}{3} \cdot \left( \frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{2|s'|} \right)$$

The results obtained, using this metric, are shown in the table below.



| Portuguese | English | Score | Correctness |
|---|---|---|---|
| iraquiana | iraqi | 0.852 | correct |
| qaida | aid | 0,867 | wrong |
| qaida | qaeda | 0,867 | correct |
| grupo | group | 0,867 | correct |
| distribuía | distribute | 0,867 | correct |
| bagdade | baghdad | 0,905 | correct |
| bagdade | bagdadi | 0,905 | wrong |
| tem | them | 0,917 | wrong |
| baghdadi | baghdad | 0,958 | wrong |
| baghdadi | bagdadi | 0,958 | correct |
| iraniana | iranian | 0,958 | correct |
| restaurante | restaurant | 0,969 | correct |
| humanitária | humanitarian | 0,972 | correct |
| abu | abu | 1 | correct |
| omar | omar | 1 | correct |
| baquba | baquba | 1 | correct |
| **Precision** | | | **75 %** |
| **Recall** | | | **70,6%** |
| **F-Measure** | | | **72,7%** |

*Table 5: test results using Jaro metric.*

The results obtained from this metric were very good. Since the Jaro distance takes into account spelling deviations, then it was already expected such great results.

However, we also obtained mistranslations. For example, the words *tem* and *them*, with the Jaro distance are said to be translations of each other. But they're not. They are very similar, differing in just one character. For these cases, there is not much we can do to avoid them. A metric measures the similarity between two words. However, very similar words in different languages can have completely different meanings.

**3.3.3. Jaro-Winkler Distance**: is an extension of the Jaro distance metric. This extension modifies the weights of pairs of strings somewhat similar which have common prefixes. This metric gives more favorable ratings to strings that match from the beginning for a set prefix length.

The Jaro-Winkler distance is given by: [3]

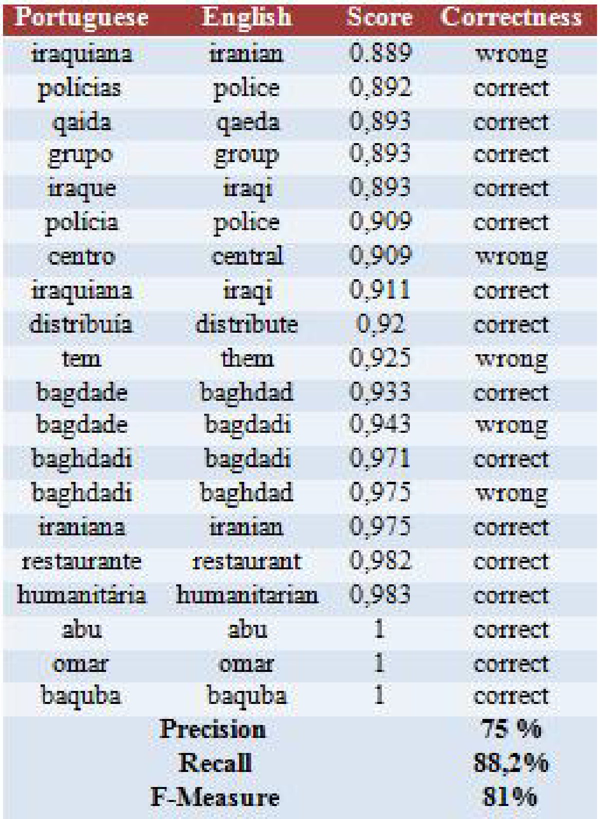
$$d_w = d_j + (\ell p(1 - d_j))$$ , where

$d_j$ is the jaro distance between two strings

$\ell$ is the length of the common prefix

$p$ is a constant scaling factor for how much the score is adjusted.

The results obtained, using this metric, are shown in the table below.



| Portuguese | English | Score | Correctness |
|---|---|---|---|
| iraquiana | iranian | 0.889 | wrong |
| polícias | police | 0,892 | correct |
| qaida | qaeda | 0,893 | correct |
| grupo | group | 0,893 | correct |
| iraque | iraqi | 0,893 | correct |
| polícia | police | 0,909 | correct |
| centro | central | 0,909 | wrong |
| iraquiana | iraqi | 0,911 | correct |
| distribuía | distribute | 0,92 | correct |
| tem | them | 0,925 | wrong |
| bagdade | baghdad | 0,933 | correct |
| bagdade | bagdadi | 0,943 | wrong |
| baghdadi | bagdadi | 0,971 | correct |
| baghdadi | baghdad | 0,975 | wrong |
| iraniana | iranian | 0,975 | correct |
| restaurante | restaurant | 0,982 | correct |
| humanitária | humanitarian | 0,983 | correct |
| abu | abu | 1 | correct |
| omar | omar | 1 | correct |
| baquba | baquba | 1 | correct |
| **Precision** | | | **75 %** |
| **Recall** | | | **88,2%** |
| **F-Measure** | | | **81%** |

*Table 6: test results using Jaro-Winkler metric.*

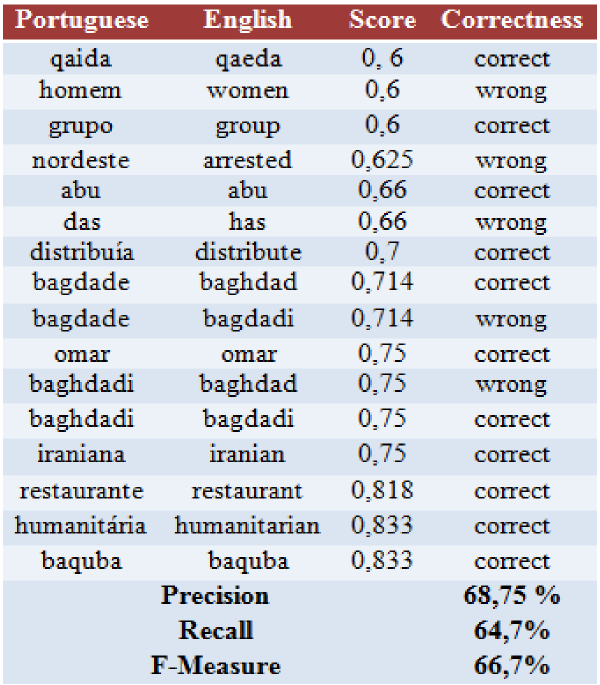
The Jaro Winkler distance processes the word prefixes, that is, it prefers words which have similar beginnings. Given that many words, in Portuguese and in English, differ only in their suffix, this metric provides extremely good results and is the most adequate for the translation of these languages. Again, we have the problems that were found in the analysis of the Jaro distance. Once more, we found that extremely similar words can have different meanings.

### 3.3.4. Longest Common Substring

The longest common substring problem is a special case of the edit distance, when substitutions are forbidden and only exact character match, insert, and delete are allowable in edit operations. The longest common substring is given by: [4]

$$C(i,j)=\begin{cases}0 & \text{if } i=M \text{ or } j=N\\ C(i+1,j+1)+1 & \text{if } x_i=y_j\\ \max\{C(i,j+1),\,C(i+1,j)\} & \text{otherwise}\end{cases}$$

The results obtained, using this metric, are shown in the table below.



| Portuguese | English | Score | Correctness |
|---|---|---|---|
| qaida | qaeda | 0, 6 | correct |
| homem | women | 0,6 | wrong |
| grupo | group | 0,6 | correct |
| nordeste | arrested | 0,625 | wrong |
| abu | abu | 0,66 | correct |
| das | has | 0,66 | wrong |
| distribuía | distribute | 0,7 | correct |
| bagdade | baghdad | 0,714 | correct |
| bagdade | bagdadi | 0,714 | wrong |
| omar | omar | 0,75 | correct |
| baghdadi | baghdad | 0,75 | wrong |
| baghdadi | bagdadi | 0,75 | correct |
| iraniana | iranian | 0,75 | correct |
| restaurante | restaurant | 0,818 | correct |
| humanitária | humanitarian | 0,833 | correct |
| baquba | baquba | 0,833 | correct |
| | **Precision** | | **68,75 %** |
| | **Recall** | | **64,7%** |
| | **F-Measure** | | **66,7%** |

*Table 7: test results using the longest common substring.*

The longest common substring tells us the maximum size that two words have in common. But this is not what we want. We want a score that allows us to choose the words that have the best results, ie which are more similar to each other. So, we developed a simple formula to obtain the score of a word. The formula is given by:

$$score(s,t)=\frac{longestCommonSubString(s,t)}{\max(length(s),length(t))}$$

With this, we are favoring the pair of words which have the longest common string most closely to the size of the biggest word.

The results obtained were good, but not that good when compared to the results obtained with the Jaro-Winkler distance metric.

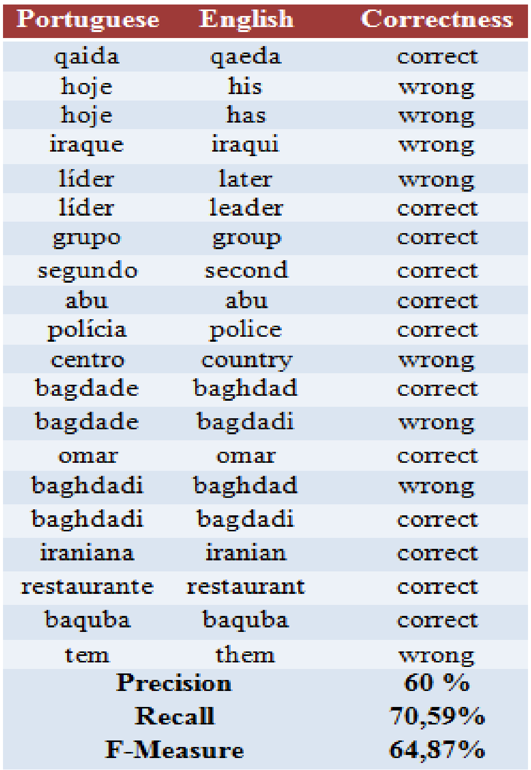
The main problem of this approach is not being able to find many pairs of words. The fact that the longest common substring does not make the treatment of word prefixes contributes to the failure to detect those pairs. On the other hand, our score formula might also contribute to the lack of results.

### 3.4. Sound Similarity

Two words are similar if they are spelled or sounded the same way. So, we tested a phonetic algorithm called Soundex [5].

Soundex is an algorithm to code surnames phonetically by reducing them to the first letter and up to three digits, where each digit is one of six consonant sounds. This reduces matching problems from different spellings.

The results obtained, using this metric, are shown in the table below.



| Portuguese | English | Correctness |
|---|---|---|
| qaida | qaeda | correct |
| hoje | his | wrong |
| hoje | has | wrong |
| iraque | iraqui | wrong |
| líder | later | wrong |
| líder | leader | correct |
| grupo | group | correct |
| segundo | second | correct |
| abu | abu | correct |
| polícia | police | correct |
| centro | country | wrong |
| bagdade | baghdad | correct |
| bagdade | bagdadi | wrong |
| omar | omar | correct |
| baghdadi | baghdad | wrong |
| baghdadi | bagdadi | correct |
| iraniana | iranian | correct |
| restaurante | restaurant | correct |
| baquba | baquba | correct |
| tem | them | wrong |
| | **Precision** | **60 %** |
| | **Recall** | **70,59%** |
| | **F-Measure** | **64,87%** |

*Table 8: test results using soundex algorithm.*

With these results, we can see two words that any of the metrics tested detected: the words (*second*, *segundo*) and (*leader*, *líder*). As we can see, their spell is different, but their sound is very similar. On the other hand, there were some mistranslations and unexpected results. The pairs (*his*, *hoje*) and (*has*, *hoje*) are a little awkward.

Soundex implements some sort of phonetic matching system. However, this system is very simple and the "letters to phoneme" mapping is a very crude model of what goes on in the English and Portuguese languages. As such, Soundex does not work for all cases. But is the only one that can find words that couldn’t be found with all the other metrics previously tested.

## 4. Algorithm Proposed

After the initial research and after testing all the metrics referred in the previous section, we propose our algorithm for the bilingual extraction of Portuguese and English words, using comparable corpora.

Our algorithm is based on the tested metrics, more particularly in the Jaro-Winkler distance and in the Soundex algorithm. We ignored all words of size less than three, because they are irrelevant and may cause errors in the extraction of bilingual lexicons.

Our algorithm is composed of three phases:

### (1) Word Pre-Process

We developed a set of more than 70 Portuguese and English rules in order to eliminate the suffixes of the words. Many of these rules were built through the consultation of Portuguese and English grammars. By doing this, we are contributing for longer word prefixes and consequently for better scores in the Jaro-Winkler distance, which will be applied next.

### (2) Similarity and Data Integration

Using the words pre-processed, we calculate the spelling similarity through Jaro-Winkler distance.

Next, we calculate the sound similarity using the phonetic algorithm Soundex. However, to prevent the algorithm from returning wrong results, we also apply the Jaro-Winkler distance, so we could be able to select words, not only similar to their sound, but also to their spell.

In this step, it is assumed that if two words are similar, then their phonetic and their spelling should not differ a lot.

Finally, we use a mediator which will be responsible for integrating the results returned from the spelling and sound similarity functions into a single file.

### (3) Result Handling

In this phase, we concentrated on eliminating possible mistranslations. We assumed that a mistranslation can occur, if a word written in Portuguese has more than one English translation. Imagine the cases *iraquiana* → *iranian* and *iraquiana* → *iraqi*. And the cases: *estou* → *am* and *sou* → am. As we can see, in the first case we have a mistranslation. By analyzing examples like this one, we assumed that if there are two translation pairs with the same Portuguese word involved, then one of those pairs might be a mistranslation. On the other hand, if we analyze the second case, the same English word can have different Portuguese translations.

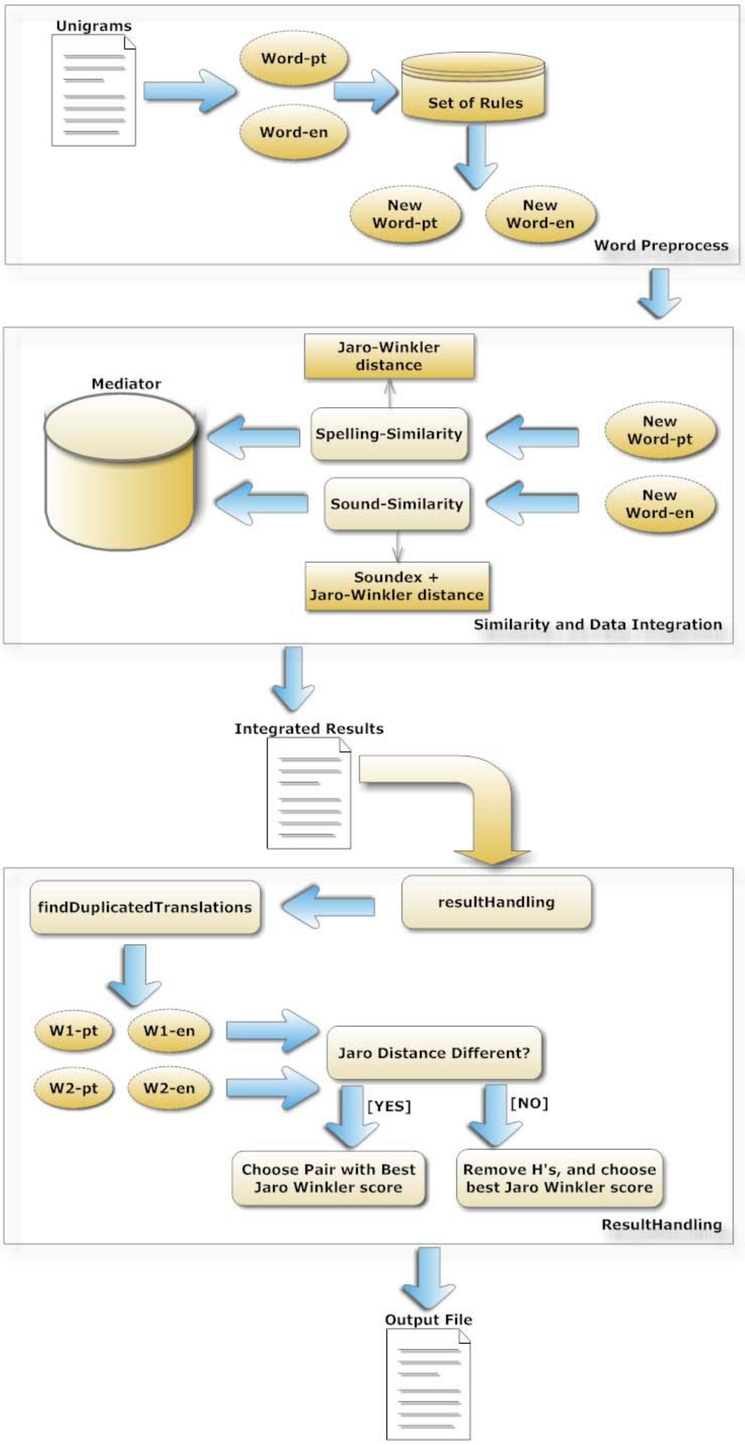


*Figure1: representation of all phases from our algorithm*

So if we detect two pairs with the same Portuguese word in it, we use our result handling function.

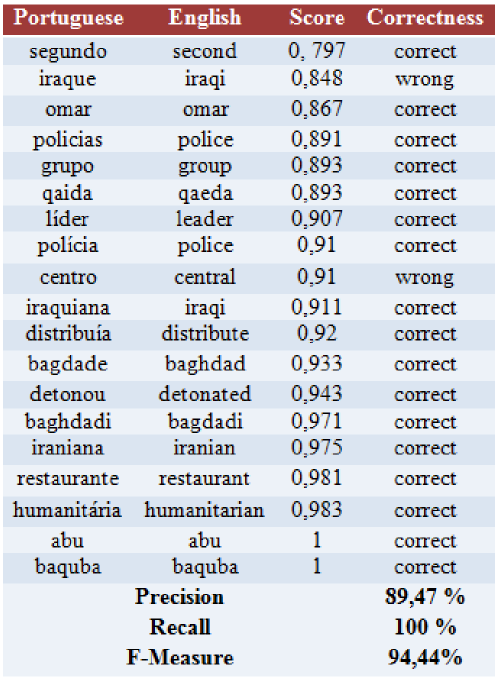
To detect which pairs were correctly translated, we used the Jaro distance as a measure of tie. Instead of scoring the pairs of words with the Jaro-Winkler, favoring their common prefixes, we used the Jaro distance to verify which pairs were more similar to each other.

However, we had a problem. What if the Jaro distance was equal? This means that the words differ in the exactly same amount of characters, becoming impossible to give different scores to the words and consequently to "choose" the best translation.

In this case, there wasn't much we could do. So, we analyzed our corpus and discovered that for some words, if we removed the "h" from them, we could get different scores using the Jaro distance. And, so we could select the correct translation. Otherwise, if the Jaro distance is different, we simply choose the pairs which have the biggest Jaro-Winkler score.

In figure 1, we show how our algorithm works for better understanding.

The results obtained, using this metric, are shown in the table below.



| Portuguese | English | Score | Correctness |
|---|---|---|---|
| segundo | second | 0, 797 | correct |
| iraque | iraqi | 0,848 | wrong |
| omar | omar | 0,867 | correct |
| policias | police | 0,891 | correct |
| grupo | group | 0,893 | correct |
| qaida | qaeda | 0,893 | correct |
| líder | leader | 0,907 | correct |
| polícia | police | 0,91 | correct |
| centro | central | 0,91 | wrong |
| iraquiana | iraqi | 0,911 | correct |
| distribuía | distribute | 0,92 | correct |
| bagdade | baghdad | 0,933 | correct |
| detonou | detonated | 0,943 | correct |
| baghdadi | bagdadi | 0,971 | correct |
| iraniana | iranian | 0,975 | correct |
| restaurante | restaurant | 0,981 | correct |
| humanitária | humanitarian | 0,983 | correct |
| abu | abu | 1 | correct |
| baquba | baquba | 1 | correct |
| | **Precision** | | **89,47 %** |
| | **Recall** | | **100 %** |
| | **F-Measure** | | **94,44%** |

*Table 9: test results using our algorithm*

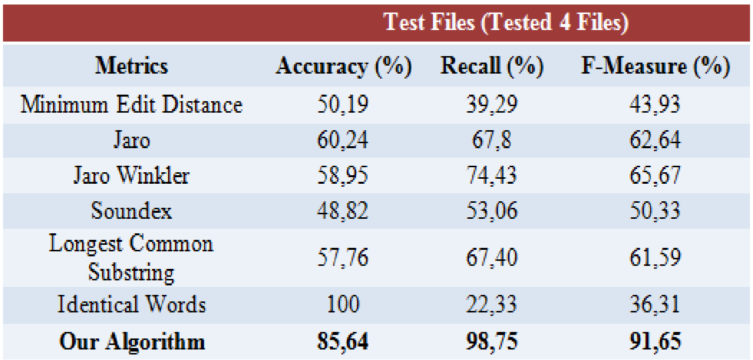
## 5 – Results

To compare the results from each technique researched against our algorithm, we chose four test corpus files and manually selected which pairs of similar words a perfect system would detect.

Then we compared the results and evaluated the performance of each metric using the accuracy, recall and f-measure.

These measures are detailed in appendix B.

The results are all summarized in table 10. However the detailed results can be found in appendix A.



| | Test Files (Tested 4 Files) | | |
|---|---|---|---|
| **Metrics** | **Accuracy (%)** | **Recall (%)** | **F-Measure (%)** |
| Minimum Edit Distance | 50,19 | 39,29 | 43,93 |
| Jaro | 60,24 | 67,8 | 62,64 |
| Jaro Winkler | 58,95 | 74,43 | 65,67 |
| Soundex | 48,82 | 53,06 | 50,33 |
| Longest Common Substring | 57,76 | 67,40 | 61,59 |
| Identical Words | 100 | 22,33 | 36,31 |
| **Our Algorithm** | **85,64** | **98,75** | **91,65** |

*Table 10: overall results*

As we can see, despite having an accuracy of 100%, identical words had the lowest scores. There aren't many words equally between the Portuguese and English, however, when they exist, most of these words have very high probability of being translations of each other. This explains why the accuracy is so high: 100%. For all words found, they were all correct. However, there weren't many words detected.

Also performed poorly, we have the minimum edit distance. This returns inaccurate results when searching for correct translation pairs.

With higher results, we have our algorithm, which is a combination of Jaro Winkler + Jaro + Soundex algorithms. The fact that we applied a word preprocessing phase, it helps in the identification of correct translation pairs. Since we are discarding suffixes, we are increasing the prefix of the words and consequently improving the results of the Jaro Winkler Distance.

## 6 – Limitations

The algorithm implemented has a major constraint: it obtains some amount of pairs of words that are not well translated. By making a word pre-process, means that we will not only be able to find more correctly words, but also be able to find many mistranslated words that neither of the metrics researched would detect.

When eliminating suffixes, the size of the new word will be smaller, allowing a match with other words that happen to have a small common prefix with the new generated word. For example:

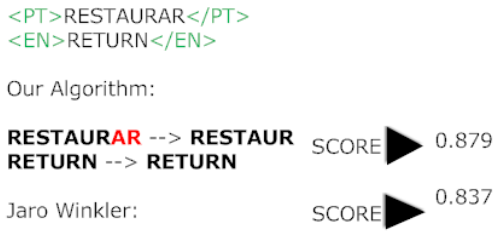


*Figure 2: an example where our algorithm fails for a threshold of 0.857. That pair of words is considered a translation using our algorithm, but a mistranslation when using the Jaro Winkler distance.*

However, some of these problems can be corrected by adding new grammar rules to the algorithm.

## 7 – Scalability

Our algorithm allows the insertion of new grammar rules that can help the searching for new pairs of words.

## 8 – Conclusions

We tried to build an algorithm allowing the extraction of bilingual lexicons between the Portuguese and English languages, mainly based on the similarity of the words, using comparable corpora.

With this attempt, we obtained several learning, such as:

1. For a big threshold, the edit distance obtains a big recall value. However, since it returns lots of mistranslated words, its accuracy is very low, being its use inadequate in the context of lexical extraction.
2. The distance of Jaro Winkler was the most appropriate metric searched. Portuguese and English words mainly differ in their suffixes. The distance of Jaro Winkler favors all words that have common prefixes. This is why we had such good results with this metric.
3. The Soundex algorithm found pairs of words that none of the metrics surveyed could. However, it has the disadvantage of detecting many mistranslations.
4. When dealing with comparable corpora, seeking translation pairs by the frequency of the words can be a dead end. With comparable corpora, nothing guarantees us that a word in Portuguese, which is very relevant, corresponds to the most relevant English word. It is not sure that a translation of a word in the Portuguese corpus can be found in the English corpus.
5. Our algorithm returned very good results. The word preprocessing phase became vital for those results. Adding rules to eliminate the words suffixes was a very good approach, because it improved the scores of the Jaro-Winkler distance.

These were our learning. All the objectives of this study were indeed met.

## 9 – Thanks

We would like to thank Professor Bruno Martins for giving us the soundex algorithm code and to clarify some doubts regarding the calculation of the frequency of a word in a document. For him a big thank you.

## 10 – References

[1] http://en.wikipedia.org/wiki/Levenshtein_distance [Levenstein distance – Wikipedia, the free encyclopedia] – last accessed on 05.01.2010

[2] http://www.dcs.shef.ac.uk/~sam/stringmetrics.html#jaro [String Similarity Metrics for Information Integration] – last accessed on 05.01.2010

[3] http://en.wikipedia.org/wiki/Jaro-Winkler_distance [Jaro-Winkler distance - Wikipedia, the free encyclopedia] – last accessed on 05.01.2010

[4] http://www.cs.princeton.edu/introcs/96optimization/ [Longest Common Substring Problem] - last accessed on 05.01.2010

[5] http://www.comp.leeds.ac.uk/matthewb/ar32/basic_soundex.htm [Soundex Algorithm] - last accessed on 05.01.2010

[6] *Gramática do Português Moderno*, José M. Castro Pinto, Manuela Neves, Maria do Céu Vieira Lopes, 1999, Plátano Editora.

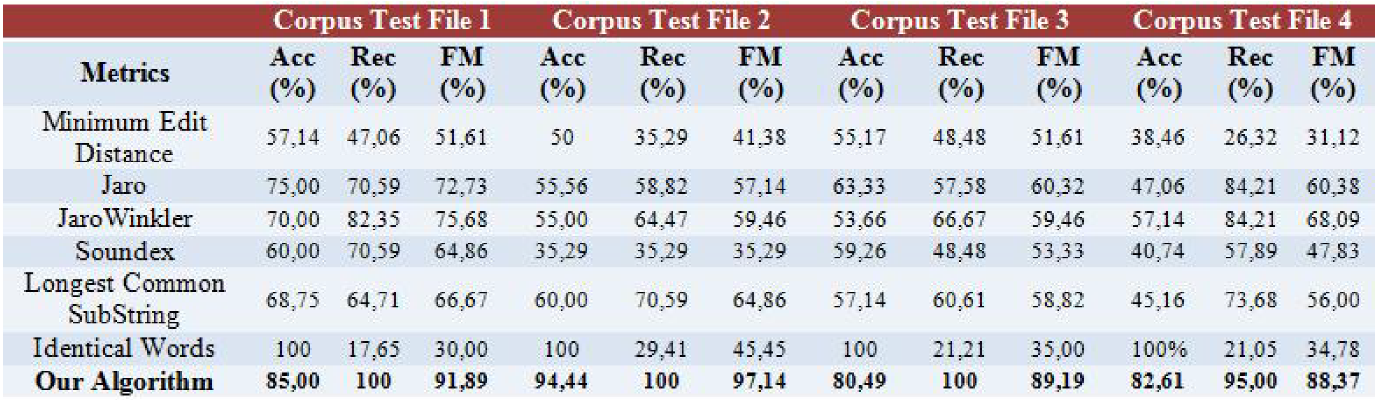
[7] http://www.learnenglish.de/grammar/suffixtext.htm [English Grammar – English Suffixes] - last accessed on 05.01.2010

[8] Gamallo P. (2008) "Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora"

## Appendix A – Detailed Results

In this section, we detail the results obtained after testing the various metrics searched against our algorithm.

The corpus test files 1, 2 and 3 have, as thematic, the politics and the corpus test file 4 has, as thematic, the economy.



| | Corpus Test File 1 | | | Corpus Test File 2 | | | Corpus Test File 3 | | | Corpus Test File 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Metrics** | **Acc (%)** | **Rec (%)** | **FM (%)** | **Acc (%)** | **Rec (%)** | **FM (%)** | **Acc (%)** | **Rec (%)** | **FM (%)** | **Acc (%)** | **Rec (%)** | **FM (%)** |
| Minimum Edit Distance | 57,14 | 47,06 | 51,61 | 50 | 35,29 | 41,38 | 55,17 | 48,48 | 51,61 | 38,46 | 26,32 | 31,12 |
| Jaro | 75,00 | 70,59 | 72,73 | 55,56 | 58,82 | 57,14 | 63,33 | 57,58 | 60,32 | 47,06 | 84,21 | 60,38 |
| JaroWinkler | 70,00 | 82,35 | 75,68 | 55,00 | 64,47 | 59,46 | 53,66 | 66,67 | 59,46 | 57,14 | 84,21 | 68,09 |
| Soundex | 60,00 | 70,59 | 64,86 | 35,29 | 35,29 | 35,29 | 59,26 | 48,48 | 53,33 | 40,74 | 57,89 | 47,83 |
| Longest Common SubString | 68,75 | 64,71 | 66,67 | 60,00 | 70,59 | 64,86 | 57,14 | 60,61 | 58,82 | 45,16 | 73,68 | 56,00 |
| Identical Words | 100 | 17,65 | 30,00 | 100 | 29,41 | 45,45 | 100 | 21,21 | 35,00 | 100% | 21,05 | 34,78 |
| **Our Algorithm** | **85,00** | **100** | **91,89** | **94,44** | **100** | **97,14** | **80,49** | **100** | **89,19** | **82,61** | **95,00** | **88,37** |

*Table 11: detailed results*

Note that Acc stands for Accuracy, Rec for Recall and FM for F-Measure. These measures are detailed in appendix B.

## Appendix B – Measures

### 1 – Accuracy

Accuracy is the fraction of the documents retrieved that are relevant to the user's information need.

$$Accuracy = \frac{\{relevantDocuments\} \cap \{retrievedDocuments\}}{\{retrievedDocuments\}}$$

### 2 - Recall

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\mathrm{Re}\,call = \frac{\{relevantDocuments\} \cap \{retrievedDocuments\}}{\{relevantDocuments\}}$$

### 3 - F-Measure

The weighted harmonic mean of accuracy and recall.

$$F - Measure = \frac{2 \times Accuracy \times \mathrm{Re}\,call}{Accuracy + \mathrm{Re}\,call}$$