Experimenting with Extracting Lexical Dictionaries from Comparable Corpora for English-Romanian language pair

Elena Irimia
Research Institute for Artificial Intelligence, Romanian Academy
Calea 13 Septembrie, No 13
elena@racai.ro

Abstract

The paper describes a tool developed in the context of the ACCURAT project (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation). The purpose of the tool is to extract bilingual lexical dictionaries (word-to-word) from comparable corpora which do not have to be aligned at any level (document, paragraph, etc.) The method implemented in this tool is introduced by (Rapp, 1999). The application basically counts word co-occurrences between unknown words in the comparable corpora and known words from a Moses extracted general domain translation table (the base lexicon). We adapted the algorithm to work with polysemous entries in the translation table, a very frequent situation which is not treated in the standard approach. We introduced other heuristics, like 1. filtration of the context vectors according to a log likelihood threshold, 2. lists of verbs (specific to each language) that can be main verbs but also auxiliary or modal verbs; 3) a cognate heuristic based on the Levenshtein Distance. The implementation can also run in multithreading mode, if the user's machine has the capacity to enable parallel execution.

1. Introduction

The task of extracting translation equivalents from bilingual corpora has been approached in different manners, according to the degree of parallelism between the source and target parts of the corpora involved. For a well sentence aligned parallel corpora one can benefit from reducing the search space for a candidate translation to the sentence dimension and external dictionaries are not required. In the case of comparable corpora, the lack of aligned segments can be compensated by external dictionaries (Rapp, 1999) or by finding meaningful bilingual anchors within the corpus based on lexico-syntactic information previously extracted from small parallel texts (Gamallo, 2007).

The word alignment of parallel corpora has been received significant scientific interest and effort starting with the seminal paper of Brown et al. (1990) and continuing with important contributions like Gale & Church (1993), Kay & Roscheisen (1993), Och, F.J. et al. (1999), etc. and many more recent approaches. They are already various free software aligners used in the industry and research, from which we mention only the famous GIZA++ (Och and Ney, 2003). Moreover, the error rate goes down to 9% in experiments made with some of these approaches (Och and Ney, 2003). By comparison, the efforts and results in extracting bilingual dictionaries from comparable corpora are much poorer. Most of the experiments are usually done on small test sets, containing words with high frequency in the corpora (>99) and the accuracy percentages are not rising above 65%.

The most popular method to extract word translations from comparable corpora, on which we based the construction of our tool, is described and used by Fung & McKeown (1997), Rapp, (1999), Chiao & Zweigenbaum, (2002). It relies on external dictionaries and is based on the following hypothesis:

word target1 is a candidate translation of word source1 if the words with which target1 co-occur within a particular window in the target corpus are translations of the words with which source1 co-occurs within the same window in the source corpus.

The translation correspondences between the words in the window are extracted from external dictionaries, being seen as *seed* word pairs. In the following table, we present, in the context of the corpus we worked on (see section 4.1), the words with which "level" tend to co-occure in the English part with their specific log-likelyhoods (ex. left column, "high level" with LL 335.0537) and the words with which a possible translation of "level", e.g. "nivelul" tend to co-occure in the Romanian part. The words in the columns are ordered so as the word in the right column on a specific line it is a possible translation of the word in the left column on the same line. (e.g.: said = anunțat, low = scăzut, mic, etc.)

level	**nivelul**
high*335.0537	ridicat*108.0321
said*111.74	anunțat*10.0774
low*110.9197	scăzut*29.3577, mic*20.6037
years*86.9735	an*16.5761
fell*83.3033	scăzut*29.3577
current*77.2435	actual*48.8756
rate*63.3928	rata*12.5533

Table 1. The words with which "nivelul" co-occurs in the Romanian corpus within a certain window (here, of length 5), listed in the right column, are translations of the words with which "level" co-occurs in the English corpus within the same window, listed in the left column.

Gamallo & Pichel (2005) used as seed expressions pairs of bilingual lexico-syntactic templates previously extracted from small samples of parallel corpus. This strategy led to a context-based approach, reducing the searching space from all the target lemmas in the corpus to all the target lemmas that appear in the same seed templates. In the improved version of the approach (Gamallo, 2007), the *precision-1* (the number of times a correct translation candidate of the test word is ranked first, divided by the number of test words) and *precision-10* (the number of correct candidates appearing in the top 10, divided by the number of test words) scores go up to 0.73 and 0.87 respectively.

In the following we will describe the algorithm implemented by our tool as introduced by Rapp (1999) and we will highlight the modifications and the adaptations we made, based on the experimental work we conducted. In Section 2 we present the original approach of Rapp, Section 3 describes our contribution to the improvement of the algorithm in the tool creation's process and Section 4 introduces the results of the experiments done on 3 types of comparable corpora.

2. Short presentation of the original approach

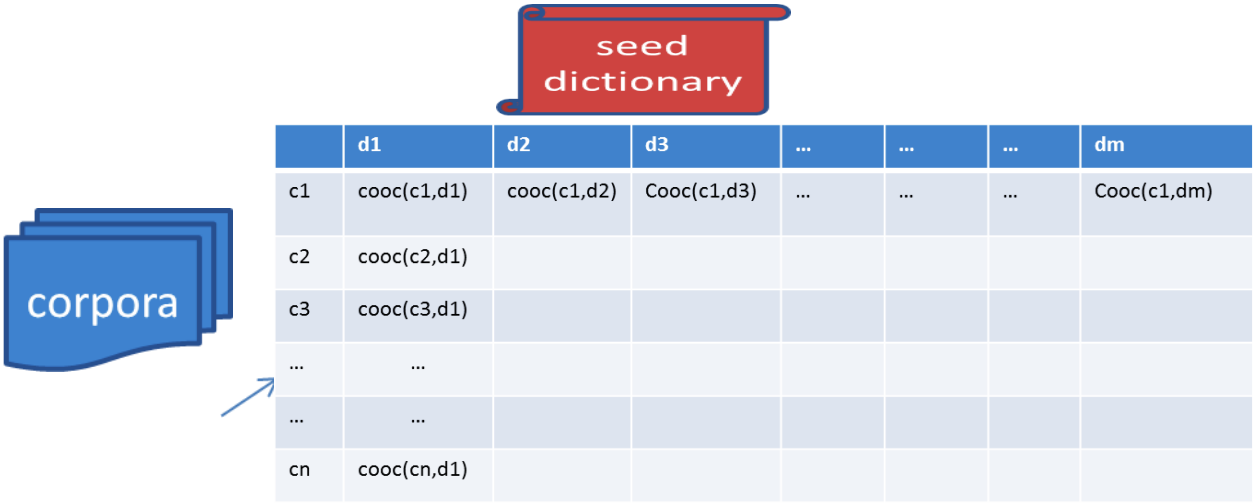
In a previous study, Rapp (1995) had already proposed a new criterion (the co-occurrence clue) for word alignment appropriate for non-parallel corpora. The assumption was that "there is a correlation between co-occurrence patterns in different languages" and he demonstrated by a study that this assumption is valid even for unrelated texts in the case of English-German language pair.

Starting from a more or less small seed dictionary and with the purpose of extending it based on a comparable corpus, a co-occurrence matrix is computed both for the source corpus and for the target corpus. Every row in the matrix corresponds to a type word in the corpus and every column corresponds to a type word in the base lexicon. For example, the intersection of a row i and a column j in the co-occurrence matrix of the source corpus contains a value sourcecooc(i,j) = frequency of common occurrence of word i and word j in a window of pre-defined size (see Figure 1 for a graphic of a generic co-occurence matrix).

The target and source corpus are lemmatized and POS-tagged and function words are not taken in consideration for translation (they are identified by their POS closed class tags: pronouns, prepositions, conjunctions, auxiliary verbs, etc.).
For any row in the source matrix, all the words with which the co-occurence frequency is above 0 are sent for translation to the seed lexicon. The unknown words (absent in the lexicon) are discarded and a vector of co-occurrences for the word correspondent to each row is computed versus the list of the translated words remained.

Experiments conducted to the need of replacing the co-occurence frequency in the co-occurence vectors by measures able to eliminate word-frequency effects and favor significant word pairs. Measures with this purpose were previously based on mutual information (Church & Hanks, 1989), conditional probabilities (Rapp, 1996), or on some standard statistical tests, such as the chi-square test or the log-likelihood ratio (Dunning, 1993). In the approach we based our tool on, the measure chosen was the log-likelihood ratio.

Finally, similarity scores are computed between all the source vectors and all the target vectors computed in the previous step, thus setting translation correspondences between the most similar source and target vectors. Different similarity scores were used in the variants of this approach; see (Gamallo, 2008) for a discussion about the efficiency of several similarity metrics combined with two weighting schemes: simple occurrences and log likelihood. Another related study was made by Laroche & Langlais (2010) which is presenting experiments around more different parameters like context, association measure, similarity measure, seed lexicon.



	d1	d2	d3	...	...	...	dm
c1	cooc(c1,d1)	cooc(c1,d2)	Cooc(c1,d3)	...	...	...	Cooc(c1,dm)
c2	cooc(c2,d1)						
c3	cooc(c3,d1)						
...	...						
...	...						
cn	cooc(cn,d1)						

Figure 1 A generic co-occurrence matrix

3. Our approach

3.1 Adaptations of Rapp's algorithm

With the aim of obtaining a dictionary similar to a translation table of the type a decoder like Moses would need to produce its translation, we decided that the lines and columns of the matrixes will be populated in our approach by word forms and not by lemmas, as in the standard approach. The option for lemma entries in the matrix was assumed also by works like (Gamallo & Pichel, 2005) and (Gamallo, 2008).

As the purpose of this tool (and of all the other tools in the ACCURAT project) was to extract from comparable corpora data that would enrich the information already available from parallel corpora, it seemed reasonable to focus (just like Rapp(1999) did) on the open class (versus closed class) words. Because in many languages, the auxiliary and modal verbs can also be main verbs, frequently basic concepts in the language (see "be" or "have" in English), and most often the POS-taggers don't discriminate correctly between the two roles, we decided to eliminate their main verb occurrences as well. For this purpose, the user is asked to provide a list of all these types with all their forms in the languages of interest (parameters: sourceamverblist, targetamverblist).

We gave the user the possibility to specify the length of the text window in which co-occurrences are counted by modifying a parameter in the configuration file. As our experiment conducted to good results for a text window of length five, this is the default value of the parameter.

Being based on word counting, the method is sensitive to the frequency of the words: the higher the frequency, the better the performance. In previous works, the evaluation protocol was conducted on frequent words, usually on those with the frequency above 100. Even in works like (Gamallo, 2008), where the evaluation was made on a list of nouns whose recall was 90% (those nouns that together come to the 90% of noun tokens in the training corpus), this corresponded to a bilingual lexicon constituted by 1,641 noun lemmas, each lemma having a token frequency>=103, for a bilingual comparable corpus of around 15 million tokens for each part. It doesn't seem too efficient to extract only a small amount of tokens from a big size corpus. Therefore, even if it causes loss of precision, the frequency threshold must be lowered when we are interested in extracting more data. In our tool, this parameter can be set by the user, according to his/her needs, but it should be above 3 (our minimal threshold) and it should take into account the corpus dimension.

As we mentioned in the previous section, the polysemy in the seed lexicon is not discussed in the standard approach. Other approaches either keep for reference only the first translation candidate in the dictionary or give different weights to the possible translations according to their frequencies in the target corpus (Morin et al., 2007).

Our seed lexicon is based on a general domain translation table automatically extracted (with GIZA++) and this is consistent with the idea that we want to improve translation data obtained from parallel corpora. But as a consequence, we deal with high ambiguity and erroneous data in the seed lexicon. In the Table 2 you can see an excerpt from the base lexicon displaying all the possible translation for the word form "creates" with their translation probabilities. Only the first three entries are exact translations of the word form "creates" while 3 of them ("instituie", "stabilește" and, in a lesser extent, „ridică" are acceptable translations in certain contexts). The two bold entries, „naștere" (birth) and „duce" (carries), may seem wrong translations learned from the training data, having a translation probability score similar to some correct translation (like "creând" or „crea"), but they also can be acceptable translations in certain contexts. We think we need to have access to all these possible translations as the semantic content of a linguistic construction is rarely expressed in another language through an identical syntactic or lexical structure. This is true especially in the case of a comparable corpus.

Our solution was to distribute the log-likelihood of a word pair (w1, w2) in the source language to all the possible translations of w2 in the target language as follows:

$$LL(w_1, w_2) = \sum_i LL(w_1, w_2) * p(w_2, t_i)$$

where $p(w_2, t_i)$ is the probability of a word w_2 to be translated with t_i and $\sum_i p(w_2, t_i) = 1$.

Every translation pair *(w_2, t_i)* is identified in the base lexicon by an unique id, making it possible to compute a similarity score across the languages.

id	*word*	*translation*	*transl. prob.*	*LL distribution*
72083	creates	creea	0.0196	LL(man,72083) =12*0.0196078 =0.2352936
72084	creates	creează	0.6862	LL(man,72084) =12*0.686275 =8.2353
72085	creates	creând	0.0196	LL(man,72085) =12*0.0196078 =0.2352936
72086	creates	**duce**	0.0196	LL(man,72086) =12*0.0196078 =0.2352936
72087	creates	instituie	0.1176	LL(man,72086) =12*0.117647 =1.411764
72088	creates	**naștere**	0.0196	LL(man,72086)

				=12*0.0196078 = 0.2352936
72089	creates	ridică	0.0392	LL(man,72089) =12*0.0392157 = 0.4705884
72090	creates	stabileşte	0.0196	LL(man,72086) =12*0.0196078 = 0.2352936

Table 2: An excerpt from the base lexicon with the possible candidate translations of the word „creates" and the distribution of LL(man, creates) = 12 according to the translation probabilities of the candidates

Previous to the LLs distribution, there is a step of LL filtering, in which all the words that occur with an LL smaller than a threshold are eliminated (the threshold is set by the ll parameter in the configuration file). This was motivated by the need to reduce the space and time computational costs and is also justified by the intuition that not all the words that occur at a specific moment together with another word are significant in the general context of our approach and the LL score is a good measure of this significance.
Following the conclusions of Gamallo's (2008) experiments, we used as a vector similarity measure the DiceMin function.

In computing the similarity scores, we did not allowed the cross-POS translation (a noun can be translated only by a noun, etc.); the user can decide if he/she allows the application to cross the boundaries between the parts of speech, through a parameter modifiable in the configuration file. Each choice has its rationales, as we know that a word is not always expressed through the same part of speech when translated in another language. On the other hand, putting all the words in the same bag increases the number of computations and the risk of error.

If the user's machine has multiple processors, the application can call a function that splits the time consuming problem of computing the vector similarities and runs it in parallel. This function is activated by the user through a "multithreading" parameter in the configuration file. To avoid overloading the memory, the application gives the user the opportunity to decide how many of the source/target vectors are loaded in the memory at a specific moment, through the "loading" parameter, activated only for "multithreading: yes"; setting this parameter to a value smaller than the matrix size can cause an important time delay, so it's in user's hands to set properly the parameters and balance advances and disadvantages according to the time constraints and according to the available memory resources.

For the proper nouns, which are more probably to be translated into a similar graphic form from a language to another, we introduced a cognate score, which is used in the computing of the similarity metric to boost the cognate candidates. This is specified in the configuration file by the parameter LD (Levenhstein Distance, the metric we based the cognate score on). This score is taken into account only if decreases under a certain threshold, which we empirically set at 0.3.

In the following, we will reproduce the configuration file we already mentioned and where the default values set for the parameters can be seen:

```
*multithreading:yes/no (default=no)
*loading:int(default=0) if the parameter's value is higher than the number of vectors in the matrix, its use becomes obsolete.
*frequency:int(default=3)
*window:int(default=5)
//5.asking for the loglikelyhood of a co-occurrence to be higher than a certain threshold, the user can reduce the space and time costs
*ll:int(default=3)
*sourceamverblist:string (default=is are be will shall may can etc.)
*targetamverblist:string (default=este sunt suntem sunteţi fi poate pot putem puteţi etc.)
*crossPOS:yes/no(default:no)
// 9.the user has to provide a list of all the open class POS labels (i.e. labels for common nouns, proper nouns, adjective, adverbs and main verbs) of the source language
*sPOSlist:string(default=nc np a r vm)
// 10.the user has to provide a list of all the open class POS labels (i.e. labels for common nouns, proper nouns, adjectives, adverbs and main verbs) of the target language
*tPOSlist:string(default=nc np a r vm)
//11.the user can decide if a cognet score (Levenshtein Distance) will be taken into account in computing the vector similarities for proper nouns
*LD:yes/no(default=no)

multithreading:yes
loading:5000
frequency:10
window:5
ll:3
sourceamverblist:am is are was were been beeing had has have be will would shall should may might must can could need
targetamverblist: este sunt eşti suntem sunteţi vei va voi vor vom veţi era eram erai eraţi fi fost pot poţi poate putem puteţi putea puteai puteam puteaţi puteau ar aţi am aş ai are avem au aveţi aveam avea aveaţi aveai aveau
crossPOS:no
sPOSlist:nc np a r vm
tPOSlist:nc np a r vm
LD:no
```

The tool is implemented in the programming language C#, under the .NET Framework 2.0. It requires the following settings to run: NET Framework 2.0., 2+ GB RAM (4 GB preferred). The application can be run as an executable file both under Windows and Linux platforms. The tool is language independent, providing that the corpus is POS-tagged according to the MULTEXT-East tag set (see http://nl.ijs.si/ME/V3/msd/html/msd.html) and that the user is introducing manually in the configuration file the list of source and target verbs concerning the parameters `sourceamverblist` and `targetamverblist`.

4.1 Experiments and results

4.1.1. Experimental setup

The base lexicon used by this tool is a word-to-word sub-part of a translation table, extracted with GIZA++ from corpora in different registers. Only the content words were kept. The translation table can be loaded as two different dictionaries EN-RO (64,613 polysemous entries) and RO-EN (66,378 polysemous entries).

Tests have been conducted on different sizes and different types/registers of comparable corpora:

1. A comparable corpora of small size representing the civil code of Romania in force until October 2011 (184,081 words) vs. the civil code of Quebec – in English (199,401 words). The corpora were manually downloaded from specific websites and we took into account the necessity to find a version of the document with diacritics for the Romanian part. The structure of the corpora is quite rigid and the noise (comprising dates or the numbers of the articles and paragraphs) was easily removed. Although we will not present detailed results here, we mention that they are not satisfactory. We assume this is due to the small size of the corpus.

2. A corpus of articles extracted at RACAI from Wikipedia: 743,194 words for Romanian, 809,137 words for English. This corpus is a strongly comparable one, with little noise (due to the fairly similar structure of the wiki pages, which facilitated the elimination of the boilerplates).

3. The corpora compiled by USFD in this project is a journalistic corpora downloaded from Google News through a heuristic based on a list of English paper titles, translated into Romanian. After the elimination of the words without content from the titles, they were used as queries into Google News and the results were downloaded for both languages. Before being released, the corpora were been cleaned for boiler plates. (For more details, see *D3.4 Report on methods for collection of comparable corpora* on the internet page of the project: http://www.accurat-project.eu/index.php?p=deliverables)

All corpora were tokenized using a library implemented in our research centre. We than checked for the presence of the diacritics and we noticed that the USFD corpora had Romanian documents which lacked those features. We used DIAC+, a tool developed at RACAI (Tufiș and Ceauşu, 2008) which automatically inserts diacritics in Romanian texts, with an error margin of 0,27% in the character accuracy.
Consequently, we checked the USFD text for repeating sentences/paragraphs and eliminated them. This reduced a lot the dimension of the USFD corpus, especially of the Romanian part.

All corpora were than lemmatized and POS-tagged using the TTL toolkit (Ion, 2007). The POS-tagging is a necessary process for selecting the content words. The output of TTL is in XML format and the annotation is compliant to the MULTEXT-East morpho-lexical specification (MSD tags, which are complex), therefore we recovered the information and put it in a simpler format (ex: *man^Nc*), keeping only the data we needed in our approach.

4.1.2. Some results

The evaluations are in progress, therefore only a small part will be presented here. We manually compiled a gold standard lexicon of around 1,500 words (common nouns, proper nouns, verbs and adjectives) from the Wikipedia corpus. In the conditions described by the default parameters in the configuration file, the precision-1 and precision-10 scores introduced earlier were computed:

POS	Precision-1	Precision-2
common nouns	0.5739	0.7381
proper nouns	0.6956	0.7336
adjectives	0.4943	0.6292
verbs	0.6620	0.8275

Table 3: P-1 and P-10 for the 1,500 test words from Wikipedia corpus

additional^af	**significant^af**
suplimentari^af 0.1268#	**importante^af** 0.0468#
general^af 0.0014#	**semnificativă^af** 0.0427#
financiare^af 0.0011#	mari^af 0.0418#
referitor^af 0.0010#	**principalele^af** 0.03902#
nouă^af 0.0008#	prezente^af 0.0367#
mari^af 0.0008#	**importantă^af** 0.0367#
indian^af 0.0007#	economice^af 0.0346#
comună^af 0.0007#	culturale^af 0.03423#
medie^af 0.0006#	**semnificative^af** 0.0339
nordică^af 0.0006#	singurele^af 0.0315#
francez^af 0.0006#	**semnificativ^af** 0.0309#
religious^af	**modern^af**
religioase^af 0.06583#	considerată^af 0.0457#
culturale^af 0.0448#	veche^af 0.0423#
politice^af 0.0412#	cunoscut^af 0.0403#
religioasă^af 0.0400#	antică^af 0.0390#
umane^af 0.0370#	roman^af 0.03790#

economice^af 0.0369#	engleză^af 0.0377#
diferite^af 0.0369#	vechi^af 0.0372#
administrativ^af 0.03474#	**modern**^af 0.0319#
sociale^af 0.0335#	latină^af 0.0314#
economic^af 0.0330#	importante^af 0.0310#
diverse^af 0.0318#	marea^af 0.0307#

Table 4: Sample of the result file for the adjective translations; the correct translations are bolded.

The experiments with the USFD corpus were very disappointing in the beginning. We realised the need for correcting some POS annotations and also to change the strategy for the LL filtration, because of the big difference in size between the two corpora (7,280,609 English words and 2,170,425 Romanian words). We decided to keep in the co-occurrence vectors only the first *n* words in descending order of their log likelihood scores. The threshold n was set experimentally to 50.

We also used the Levenshtein Distance for all the POS analysed to boost the scores for the translations graphically more similar with the word to be translated. This boost is done after all the similarity scores between a certain source word and all the target words are computed. The threshold to which the words were considered cognates were a LD<0.3 and the boost meant a multiplication with 10 of the similarity score. All the scores that resulted above 1 were reduced to 0.99.

We also felt the need for introducing two different frequency thresholds for the two corpora, to compensate the difference in size. The values of the frequencies established after more experiments were 100 for the source words (English) and 20 for the target words (Romanian).

After all these heuristics, the results become more reasonable, but still not rising to the performances on the Wikipedia corpus. We explain that but the serious difference in the degree of comparability between the corpora.

Because of the time constraints (the final and cleaner version of the USFD corpora was made available shortly before the deadline for this paper) we focused only on three POS: common nouns, adjectives and verbs. We constructed for each POS a gold-standard dictionary with 100 entries and Precision-1 and Precision-10 scores were computed:

POS	Precision-1	Precision-10
common nouns	0.2909	0.5454
adjectives	0.3663	0.5049
verbs	0.24	0.48

Table 5: P-1 and P-10 for the 300 test words from USFD corpora

The effect of introducing the cognate test for all the POS was important for many of the good results, producing more forms of the same lemma as possible translations, which is consistent with the reach morphology of Romanian and is very useful in a dictionary:

ministers^nc|*ministru^nc ministrul^nc miniştrilor^nc*
fund^nc|*fondului^nc fondul^nc fond^nc*
sector^nc|*sector^nc sectorul^nc sectorului^nc*

republican^af|*republican^af republicani^af republicană^af*
national^af|*naţional^af naţională^af naţionale^a*
german^af|*german^af germană^af germane^af germani^af*

considered^vm|*considerat^vm consideră^vm considera^vm consider^vm considerând^vm*
continue^vm|*continua^vm continuă^vm continue^vm continuat^vm*
confirm^vm|*confirmat^vm confirmă^vm confirma^vm confirmată^v*

This phenomenon occurred for around 46% of the correct translated nouns, 39% of the correct translate adjectives and 29% of the correct translated verbs.

For some translations in which the cognate test didn't interfered, multiple solutions could be seen also:

policies^nc|*plan^nc program^nc planul^nc măsurilor^nc măsuri^nc*
debts^nc|*datoriile^nc datoriilor^nc*

former^af|*fostul^af fostului^a*
black^af|*negru^af neagră^af*
last^af|*trecut^af fostul^af recent^af*

played^vm|*juca^vm jucat^vm*
earned^vm|*câştigat^vm obţinut^vm*
die^vm|*murit^vm mor^vm muri^vm moară^vm*

5. Conclusions

We created a tool destined to extract bilingual word-to-word lexicons from comparable corpora. Based on a well-known approach (Rapp, 1999) we intended to extend it to deal with polysemy, so that we can use automatically extracted translation tables as seed dictionaries. We also proposed a filtration of the co-occurrence vectors according to the log likelihood score, starting from the idea that this score is a good measure for the significance of two words occurring together. The tool can be also used in multithreading mode if the user's machine has multiple processors.

From the three types of corpora we experimented with, only one (No.2) showed good and really usable results. This is coming from the strong comparability of the corpora (Wikipedia articles are quite similar, with some in one language being poorer in content than in the other language). We will keep working on the corpus No. 1, by adjusting the parameters in the configuration file and on the corpus No.3 by experimenting with the LL score

filtration. We also need to evaluate how many new words (which are not part of the seed dictionary) are translated through our method.

6. Acknowledgements

Research reported in this paper has been supported by the ACCURAT project within the EU 7th Framework Pogramme (FP7/2007-2013), grant agreement no. 248347.

References

Brown, P.; Cocke, J.; Della Pietra, S. A.; Della Pietra, V. J.; Jelinek, F.; Lafferty, J. D.; Mercer, R. L.; Rossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2), 79-85.

Chiao, Y.-C., Zweigenbaum, P. (2003) Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of Coling 2002*, Taipei, Taiwan, 26-30 August 2002

Church, K. W., Hanks, P. (1989). Word association norms, mutual information, and lexicography. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. Vancouver, British Columbia, 76-83.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.

Fung, P., McKeown, K. (1997) Finding terminology translations from non-parallel corpora. *Proceedings of the Fifth workshop on Very Large Corpora*. ed. Joe Zhou and Kenneth Church, 18 August 1997, Tsinghua University, Beijing, China, 20 August 1997, Hong Kong University of Science and Technology, Hong Kong; pp.192-202.

Gamallo P., Pichel, J.R. (2005). An Approach to Acquire Word Translations from NonParallel Texts, *Lecture Notes in Computer Science*, vol. 3808. SpringerVerlag.

Gamallo, P. (2007). Learning bilingual lexicons from comparable english and spanish corpora. In *Machine Translation SUMMIT XI*, Copenhagen, Denmark.

Gamallo P. (2008) Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. *Proceedings of LREC 2008 Workshop on Comparable Corpora*, Marrakech, Marroco, pp. 19-26. ISBN: 2-9517408-4-0.

Gale, W. A., Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(3), 75-102.

Kay, M., Roscheisen, M. (1993). Text-Translation Alignment. *Computational Linguistics*, 19(1), 121-142.

Laroche A., Langlais P.: Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of COLING 2010*: 617-625.

Morin E., Daille B., Takeuchi K., Kageura K. (2007). Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (ACL'07), pp. 664-671.

Och, F.J. and Ney, H. (2003) A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003

Och, F. J., Tillmann, C., Ney, H. (1999). Improved alignment models for statistical machine translation. *Joint SIGDAT conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Proceedings ed. Pascale Fung and Joe Zhou*, 21-22 June 1999, University of Maryland, College Park, MD, USA; pp.20-28.

Rapp, R. (1995). Identifying word translations in nonparallel texts. In: *Proceedings of the 33rd Meeting of the Association for Computational Linguistics. Cambridge*, Massachusetts, 320-322.

Rapp, R. (1996). Die Berechnung von Assoziationen. Hildesheim: Olms.

Rapp, R. (1999) Automatic identification of word translations from unrelated English and German corpora. *ACL-1999: 37th Annual Meeting of the Association for Computational Linguistics. Proceedings of the conference*, 20-26 June 1999, University of Maryland, College Park, Maryland, USA; pp.519-526.

Tufiş, D, Ceauşu A. (2008). DIAC+: A Professional Diacritics Recovering System, in *Proceedings of LREC 2008 (Language Resources and Evaluation Conference)*, May 26 - June 1, Marakkech, Morocco. ELRA - European Language Resources Association. ISBN: 2-9517408