QAlign: A New Method for Bilingual Lexicon Extraction from Comparable Corpora

Amir Hazem and Emmanuel Morin

Laboratore d'Informatique de Nantes-Atlantique (LINA)
Université de Nantes, 44322 Nantes Cedex 3, France
{Amir.Hazem,Emmanuel.Morin}@univ-nantes.fr

Abstract. In this paper, we present a new way of looking at the problem of bilingual lexicon extraction from comparable corpora, mainly inspired from information retrieval (IR) domain and more specifically, from question-answering systems (QAS). By analogy to QAS, we consider a word to be translated as a part of a question extracted from a source language, and we try to find out the correct translation assuming that it is contained in the correct answer of that question extracted from the target language. The methods traditionally dedicated to the task of bilingual lexicon extraction from comparable corpora tend to represent the whole contexts of a word in a single vector and thus, give a general representation of all its contexts. We believe that a local representation of the contexts of a word, given by a window that corresponds to the query, is more appropriate as we give more importance to local information that could be swallowed up in the volume if represented and treated in a single whole context vector. We show that the empirical results obtained are competitive with the standard approach traditionally dedicated to this task.

Keywords: Comparable corpora, bilingual lexicon extraction.

1 Introduction

The use of comparable corpora for the task of bilingual lexicon extraction has attracted great interest since the beginning of 1990. Introduced by [25] he assumes that algorithms for sentence and word alignment from parallel texts should also work for non parallel and even unrelated texts. Comparable corpora offer a great alternative to the inconvenience of parallel corpora. Parallel corpora are not always available and are also difficult to collect especially for language pairs not involving English and for specific domains, despite many previous efforts in compiling parallel corpora (Church & Mercer, 1993; Armstrong & Thompson, 1995). According to Rapp [25] : *The availability of a large enough parallel corpus in a specific field and for a given pair of languages will always be the exception, not the rule.* Since then, many investigations and a number of studies have emerged, [10,11,12,24,26,2,7,14,23,21, among others]. All these works are based on a general representation of the contexts of a given word by collecting all its co occurrences in a single large vector. We want to give particular attention to each context as it represents a specific idea that can be lost if treated in a whole context vector. QAS systems alleviate this drawback and offer a suitable environment for our

A. Gelbukh (Ed.): CICLing 2012, Part II, LNCS 7182, pp. 83–96, 2012.

task. Basically, the aim of a question answering system is to find the correct answer to a given question. The main idea of such a QAS is to consider segments or paragraphs of documents that share several words with a given question and then order them according to a similarity measure [15]. Those n best segments are most likely to provide the correct answer. Complex systems will not only use the words of the question but also synonyms or other semantically related words. More sophisticated systems will reformulate the question and so on [28][19][20][22][16]. In a multilingual context, the question is first translated and then the same treatments are applied as stated previously. In our case, we want to push QAS systems a step further by considering the bilingual lexicon extraction from comparable corpora as a question answering system, where the question is one of the contexts of the word to be translated, and the best answer should be the one containing the correct translation in the target language. In this case and for a given word we have as many questions as this word occur. This can be a problem if a word has a high frequency. We obviously cannot consider all the contexts of such a word, this is not our aim. On the contrary we will consider the n best contexts which is one part of the problem that we have to deal with. The remainder of this paper is organised as follows. Section 2 presents the standard approach based on lexical context vectors dedicated to word alignment from comparable corpora. Section 3 describes our Q-Align approach that can be viewed as a question answering system for alignment. Section 4 describes the different linguistic resources used in our experiments. Section 5 evaluates the contribution of the standard and Q-Align approaches on the quality of bilingual terminology extraction through different experiments. Section 6 presents our discussion and finally, Section 7 presents our conclusions and some perspectives.

2 Standard Approach

The main work in bilingual lexicon extraction from comparable corpora is based on lexical context analysis and relies on the simple observation that a word and its translation tend to appear in the same lexical contexts. The basis of this observation consists in the identification of first-order affinities for each source and target language: "*First-order affinities describe what other words are likely to be found in the immediate vicinity of a given word*" [17, p. 279]. These affinities can be represented by context vectors, and each vector element represents a word which occurs within the window of the word to be translated (for instance a seven-word window approximates syntactical dependencies). The implementation of this approach can be carried out by applying the four following steps [25,13]:

Context Characterisation. All the lexical units in the context of each lexical unit i are collected, and their frequency in a window of n words around i extracted. For each lexical unit i of the source and the target languages, we obtain a context vector $\mathbf{i}$ where each entry, $\mathbf{i}_j$, of the vector is given by a function of the co-occurrences of units j and i. Usually, association measures such as the mutual information [9] or the log-likelihood [8] are used to define vector entries.

Vector Transfer. The lexical units of the context vector $\mathbf{i}$ are translated using a bilingual dictionary. Whenever the bilingual dictionary provides several translations for a lexical unit, all the entries are considered but weighted according to their frequency in the target language. Lexical units with no entry in the dictionary are discarded.

Target Language Vector Matching. A similarity measure, $\mathrm{sim}(\overline{\mathbf{i}}, \mathbf{t})$, is used to score each lexical unit, t, in the target language with respect to the translated context vector, $\overline{\mathbf{i}}$. Usual measures of vector similarity include the cosine similarity [27] or the weighted jaccard index (WJ) [18] for instance.

Candidate Translation. The candidate translations of a lexical unit are the target lexical units ranked following the similarity score. The translation of the lexical units of the context vectors, which depends on the coverage of the bilingual dictionary vis-à-vis the corpus, is an important step of the standard approach, as more elements of the context vector are translated, the context vector will be more discriminating in selecting translations in the target language. This drawback can be partially circumvented by combining a general bilingual dictionary with a specialised bilingual dictionary or a multilingual thesaurus [3,7]. Moreover, this approach is sensitive to the choice of parameters such as the size of the context, the choice of the association and similarity measures. The most complete study about the influence of these parameters on the quality of bilingual alignment has been carried out in [21]. Another approach has been proposed to avoid the insufficient coverage of the bilingual dictionary required for the translation step of the standard approach [7,5]. The basic intuition of this approach is that words that have the same meaning will share the same environments. Here, the approach consists in the identification of "Second-order affinities" for the source language: "*Second-order affinities show which words share the same environments. Words sharing second-order affinities need never appear together themselves, but their environments are similar*" [17, p. 280]. For a word to be translated its affinities can be extracted through distributional techniques. In this case, the translation of a word consists of the translation of similar words. Since this approach is sensitive to the size of the comparable corpus, this study focuses on the standard approach.

3 Q-Align Approach

The Q-Align approach is described in three steps as follows :

3.1 Collecting the Queries

The first step of the Q-Align approach, is to collect the set of all the windows (queries) in which a word to be translated appears. The size of this set corresponds to the frequency of the candidate. We have to deal with two parameters, the first one is the size of each query, it can be seen as the window surrounding the word to be translated as usually done in the state of the art. Let us call this parameter w_q. For instance, let us take **replica** as the word to translate, if $w_q = 5$ this means that there are two words on the

Table 1. English query of the word **replica**

$detail_V$	$paintings_S$	**replica**$_S$	$line_V$	$separate_J$

left of **replica** and two words on its right. After the POS-Tagging and filtering process, we obtain the resulting query for the word **replica** as shown in Table 1.

The second parameter is the number of queries we need for our task. We start from the assumption that not all contexts are useful when trying to find the correct translation. On the contrary, some of them are useless and can be considered as noise. Following this principle, we believe that a good choice of a context maximises the chances of matching the correct translation. Several ways can be followed to deal with this parameter in order to find the best tuning. As we wanted to focus on the comparison between the standard and Q-Align approaches in term on context characterisation, we did not investigate the different possibilities of choosing the number of queries, which is on it self a great matter of interest for future work. we fixed this parameter empirically. The choice of the n best queries was merely done following equation 1 :

$$Score(query_n) = \sum_{i=1}^{w_q - 1} freq(word_i) \tag{1}$$

After applying the calculation of the score for all the queries, we sorted in a decremental order the n queries according to $Score(query_n)$.

3.2 Translation of Queries

Each collected query has to be translated into the target language, if we use the previous example of the word **replica**, and if we consider French as the target language, we obtain the corresponding translation query in Table 2 :

Table 2. Representation of the English query of the word **replica** and its translation into French

Word	Translation
$detail_V$	$désigner_V$
$paintings_S$	$paintures_S$
replica$_S$	**Unknown**$_S$
$line_V$	$marquer_V$
$separate_J$	$indépendant_J$

The translated query that will be used in the target language is given in Table 3 :

Table 3. Translated query of the word **replica**

$désigner_V\ paintures_S\ marquer_V\ indépendant_J$

It is worth noting that the words of the query are translated using a bilingual dictionary while preserving the POS-Tagging relation of each translation pair. When several translations for a word are given, we consider the one with the highest frequency in the target language. Words with no entry in the dictionary are discarded.

3.3 Extraction of the Translation Candidates

To select a translation candidate, we use the compactness [28] as similarity measure. The principle of compactness in QAS is to measure a similarity between a question and a given segment. A segment can be : a sentence, a paragraph or a document. In our case, and by analogy, we measure the compactness between a translated query and a given segment of a given document in the target corpus. The final compactness $Compact_{All}(\bar{w_x})$ of $\bar{w_x}$ is simply the sum of its compactness according to all translated queries, as given by the following equation :

$$Compact_{All}(\bar{w_x}) = \sum_{i \in nbQuery} Compact(\bar{w_x})_i \tag{2}$$

All the documents of the target language are divided into segments. We investigate each segment to find out if it contains the correct translation. We need to fix the size of the segments. Let us denote w_{seg} as the size of a given segment corresponding to the number of words that belongs to this segment. For a given translated query and a given segment, the compactness of $\bar{w_x}$ for a segment s is given by :

$$Compact_s(\bar{w_x}) = \frac{1}{|WQ|} \sum_{i=\in WQ} Contrib(w_i)_{\bar{w_x}} \tag{3}$$

where $Contrib(w_i)_{\bar{w_x}}$ is the contribution of each word of the query. Let us give an example to illustrate how to compute the contribution and the compactness. We denote QR as the set of words of the translated query as shown in table 4, with $w_q = 5$ and w_i a word of the given query. In the example $QR = \{w_1, w_2, w_3, w_4\}$ and $\mathbf{Cand}_S$ is the word to be translated .

Table 4. English query of the word to be translated

w_1	w_2	$\mathbf{Cand}_S$	w_3	w_4

Let us consider also, a segment with $w_{seg} = 8$. Each word of the segment which is not part of the question is considered as a translation candidate, we can take $\bar{w_x}$ as a candidate :

We compute the contribution of each word $w_i \in QR$ surrounding $\bar{w_x}$ following this equation:

$$Contrib(w_i)_{\bar{w_x}} = \frac{|Z|}{D+1} \tag{4}$$

Table 5. Representation of a given segment

w_1			w_2	$\bar{w_x}$		w_3	$\bar{w_4}$	w_4
-4	-3	-2	-1	0	1	2	3	4

Where : $D = distance(w_i, \bar{w_x}) = |pos(w_i) - pos(\bar{w_x})|$

$pos(w_i)$ is the position of w_i in a given segment, for instance, in Table 5, $pos(w_1) = -4$.

$Z = \{Y \setminus distance(Y, \bar{w_x}) < D \quad and \quad Y \in QR\} \cup \{\bar{w_x}\}$

For example the contribution of w_1 is given by :

$$Contrib(w_1)_{\bar{w_x}} = \frac{2+1}{4+1} = \frac{3}{5} \tag{5}$$

We wanted to consider differently the words of a given translated query, one way was to weight the contribution of a word by its Inverse Segment Frequency (ISF) by analogy to Inverse Document Frequency (IDF)[16][15] , assuming that words with high ISF should be more important. This can be seen in equation 6 :

$$Compact_s(\bar{w_x}) = \frac{1}{|WQ|} \sum_{i=\in WQ} ISF(w_i) \times Contrib(w_i)_{\bar{w_x}} \tag{6}$$

We have given above the compactness of a word computed from a given segment. As there is thousands of segments in a corpus, we chose the maximum compactness of a given word according to equation 7. Some other alternatives have been explored as the mean compactness or the sum but no significant differences or improvements have been noticed.

$$Compact(\bar{w_x}) = \max(Compact_s(\bar{w_x})) \tag{7}$$

Starting from the intuition that the translation of a rare word in a source language should also be rare in the target language following the principle of comparable corpora, we weighted the final compactness for rare words by the ISF. This is represented in equation 8 :

$$Compact_{All}(\bar{w_x}) = \sum_{i \in nbQuery} ISF(\bar{w_x}) \times Compact(\bar{w_x})_i \tag{8}$$

No other alternative except the weighted sum showed a significant improvements, but more investigations have to be conducted especially on the choice of the good queries. This represents our next challenge.

4 Linguistic Resources

The experiments have been conducted on two different French-English corpora: a specialised corpus from the medical domain within the sub-domain of 'breast cancer' and a general corpus from newspapers 'LeMonde/New-York Times '. Due to the small size

of the specialised corpus we wanted to conduct additional experiments on a large corpus to have a better idea of the behaviour of our approach. Both corpora have been normalised through the following linguistic pre-processing steps: tokenisation, part-of-speech tagging, and lemmatisation. The function words have been removed and the words occurring less than twice (i.e. hapax) in the French and the English parts have been discarded.

4.1 Specialized Corpus

We have selected the documents from the Elsevier website[1] in order to obtain a French-English specialised comparable corpus. We have automatically selected the documents published between 2001 and 2008 where the title or the keywords contain the term 'cancer du sein' in French and 'breast cancer' in English. We collected 130 documents in French and 118 in English and about 530,000 words for each language. The comparable corpus comprised about 7,400 distinct words in French and 8,200 in English. In bilingual terminology extraction from specialised comparable corpora, the terminology reference list required to evaluate the performance of the alignment programs is often composed of 100 single-word terms (SWTs) (180 SWTs in [6], 95 SWTs in [2], and 100 SWTs in [5]). To build our reference list, we selected 400 French/English SWTs from the UMLS[2] meta-thesaurus and the *Grand dictionnaire terminologique*[3]. We kept only the French/English pair of SWTs which occur more than five times in each part of the comparable corpus. As a result of filtering, 122 French/English SWTs were extracted.

4.2 General Corpus

We chose newspapers as they offer a large amount of data. We selected the documents from the French newspaper 'Le Monde' and the English newspaper 'The New-York Times '. We automatically selected the documents published between 2004 and 2007 and obtained 5 million words for each language. The comparable corpus comprised about 70,400 distinct words in French and 80,200 in English. The terminology reference list is much more consequential and contains 1004 SWTs, it has been extracted from ELRA-M0033. We divided this list into 8 sub-lists according to word frequency as presented in Table 6 :

4.3 Bilingual Dictionary

The French-English bilingual dictionary required for the translation phase was the ELRA-M0033 dictionary. It contains, after linguistic pre-processing steps, 32,000 English single words belonging to the general language with an average of 1.6 translations per entry.

[1] www.elsevier.com
[2] www.nlm.nih.gov/research/umls
[3] www.granddictionnaire.com/

Table 6. Representation of each evaluation list

Name List	Interval	$\#occ$
$List\_sup\_1000$	$\#occ > 1000$	4
$List\_500\_1000$	[500, 1000[	20
$List\_100\_500$	[100, 500[	180
$List\_50\_100$	[50, 100[	200
$List\_10\_50$	[10, 50[	400
$List\_2\_10$	[2, 10[	200

5 Experiments and Results

In this section, we first give the parameters of the standard and Q-Align approaches, than we present the results conducted on the two corpora presented above: "Breast cancer" and "LeMonde/New-YorkTimes".

5.1 Experimental Setup

Three major parameters need to be set to the standard approach, namely the similarity measure, the association measure defining the entry vectors and the size of the window used to build the context vectors. Laroche and Langlais [21] carried out a complete study of the influence of these parameters on the quality of bilingual alignment. As a similarity measure, we chose to use the weighted jaccard index [18]. The entries of the context vectors were determined by the log-likelihood [8], and we used a seven-word window since it approximates syntactic dependencies. Other combinations of parameters were assessed but the previous parameters turned out to give the best performance. For the Q-Align approach we also used a seven-word window that corresponds to the query length. The size of segments in the target language was fixed to one hundred words even if several combinations were assessed. This size gave the best performance on a fixed length query of seven words. The choice of one hundred as the length of a segment is due to the fact that it is more or less the length of a paragraph.

5.2 Results

To evaluate the performance of our approach, we used the standard approach (SA) proposed by [26] as a baseline. The accuracy is given in percentage in all the graphics.

Evaluation on the Breast Cancer Corpus

We investigate the performance of the Standard and Q-Align approaches on the breast cancer corpus, using the evaluation list of 122 words.

We can see in Figure 1 that Q-Align approach always outperforms the standard approach for all values of k. The accuracy at the top 20 for the standard approach is 54.91% while Q-Align approach gives 60.62%. The Q-Align model can be considered as a competitive approach according to its results as shown in Figure 1 for the breast cancer corpus.

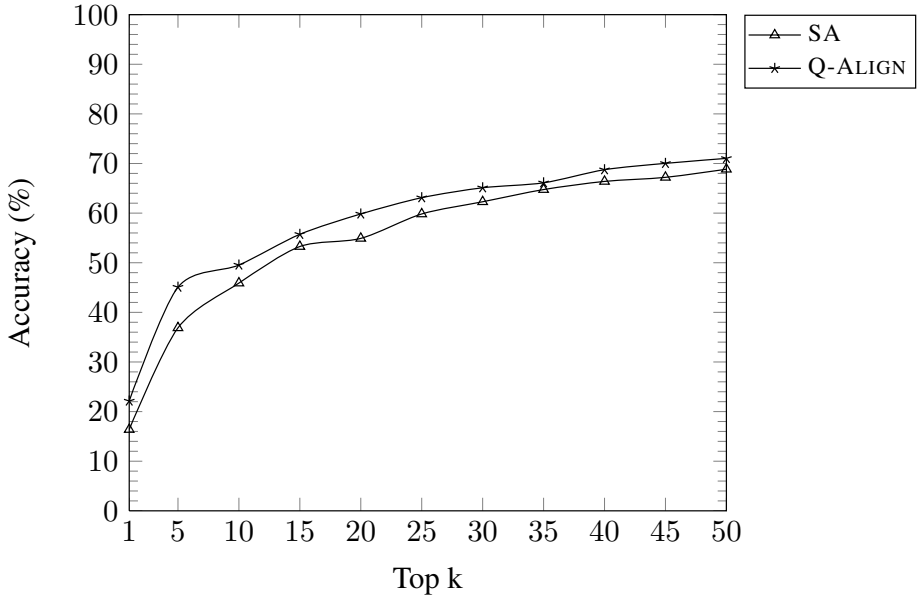

Fig. 1. Accuracy at top k for the breast cancer corpus (SA vs Q-Align)

Evaluation on the LeMonde/New-YorkTimes Corpus

We then investigate the performance of the Standard and Q-Align approaches on LeMonde/New-YorkTimes corpus, using an evaluation list of 1004 words.

Let us see in Figures 3 and 4 the details of the ranges of frequency into which Q-Align and standard approaches failed.

Figure 3 and 4 show that both approaches are sensitive to the variations of word's frequencies. It seems that the Q-Align approach is slightly less efficient for rare words with frequencies less than 50 while the standard approach (SA) is slightly better. Similarly for very frequent words with frequencies higher than 500 the standard approach outperforms Q-Align approach except for top 1, 5 and 10. The main gap between SA and Q-Align in term of accuracy can be seen for the lists where the words frequencies are between 50 and 500. Due to the small list of words with frequencies higher than 1000, we cannot give an appropriate conclusion for both approaches as the number of words in this list is equal to 4. The main reason for the weakness of the Q-Align approach in a general corpus is probably the lack of markers or seed words that are more present in a specialised corpus. In the light of these results more investigations have to be conducted on general corpus to improve the performance of the Q-Align approach.

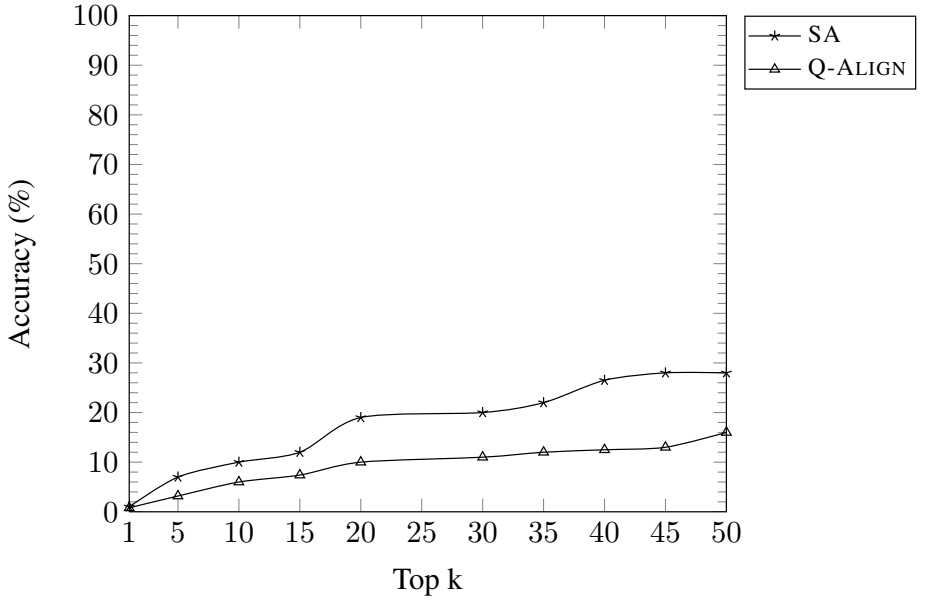

Fig. 2. Accuracy at top k for LeMonde/NewYorkTimes (SA vs Q-Align)

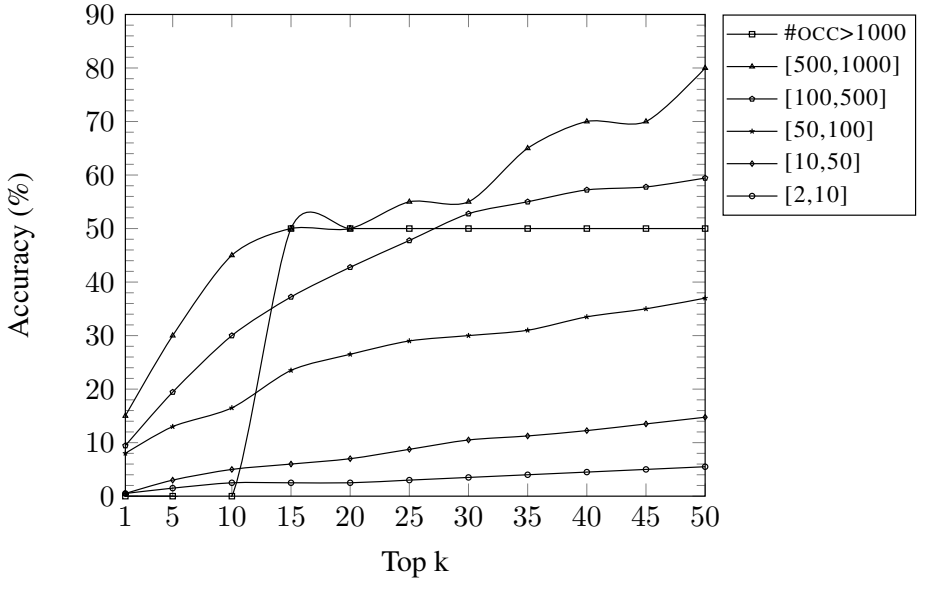

Fig. 3. Accuracy at top k for LeMonde/NewYorkTimes (Standard Approach)

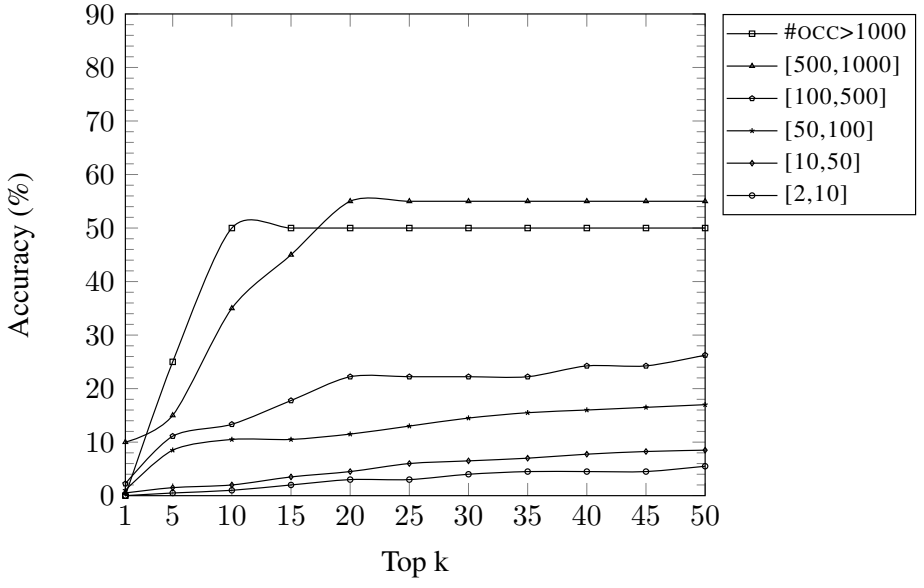

Fig. 4. Accuracy at top k for LeMonde/NewYorkTimes (Q-Align Approach)

6 Discussion

The aim of this work was not to try to find the best results by looking for the best tuning of each method (SA, Q-Align). Here, the main interest was to show another way of looking at the task of bilingual lexicon extraction from comparable corpora by looking at local information captured in a given segment while the standard approach looks at global information captured in the whole corpus. The Q-Align approach imitates QAS systems by choosing a query in the source language and tries to find an answer in a particular segment of the target language which contains the correct translation. This process was done by comparing the translated query with segments that we consider more or less close to paragraphs in the target language, assuming that if a given paragraph shares some words with the translated query then, there is more chance that this paragraph contains the correct translation.

It is important to say that Q-Align is a naive approach. The process of looking for the right translation does not take into account any linguistic or semantic information, it is just based on words that are in common between a query and a segment. Surprisingly, this naive approach (Q-Align) outperforms the standard approach (SA) for each top on the specialised corpus. Thus, there is much to do to improve the Q-Align approach while no semantic or linguistic information was taken into account. Q-Align can be considered as promising for future work. Many improvements need to be done. QAS systems are a great source of inspiration for it.

It would be interesting to merge both approaches to see whether there is a complementarity or not between them and if obviously there is an increase in accuracy.

We can reflect upon a double interest by using both approaches. The first one is given by the standard approach (SA) as it gives a global view and a global representation of information. The second one is given by the Q-Align approach which gives a local view and a local representation of information. We can imagine that both local and global information could be useful taken together to improve the representation of information and thus to obtain increased accuracy.

In the Q-Align approach one drawback is probably the choice of queries. Indeed, not all the queries are useful when trying to find the right translation. Some of them bring more confusion then good information. We can consider these queries as noise, because if taken, they could lead us to wrong translations. Following this principle, we believe that a good choice of a query maximises the chances of matching the right translation. Several alternatives to our arbitrary choice of the number of queries can be applied in order to find the best tuning. In this paper we did not explore the way of choosing the number of queries. This question remains opened and represents one of the unanswered questions. It is for us one of the next challenges.

We must add that our research concentrated on specialised corpora as it is our main center of interest. We were however curious to see the behaviour of our approach on a general corpus and the results were as expected. At the moment, Q-Align approach is more appropriate to a specific rather than to a general domain as all our efforts were conducted in this way. This performance can be explained by the specificity of specialised corpora such as the medical domain for instance. which contains strong markers that are for the greater part technical words or words specific to the domain. These markers that we can see as seed words are very useful for the Q-Align approach. The results obtained clearly point to one conclusion which is that Q-Align approach is more appropriate for corpora of specialised domains while for general corpora it remains unstable due the the lack of specific markers. More efforts on general domain data have to be made to adapt Q-Align approach to general domain corpora.

7 Conclusion

We have presented a novel way of looking at the problem of bilingual lexicon extraction from comparable corpora based on the principle of question answering systems. We explored two different corpora, the first concerned a corpus of medical domain (Brest Cancer) and the second concerned the corpus of newspapers (LeMonde/NewYorkTimes). Regarding the empirical results of our proposition, performances were better than the baseline proposed by [26] on specialised corpus. While the standard approach remained more robust in a general domain corpora. Further research is certainly needed but our current findings support the idea that local information has its importance and should not be neglected for the task of bilingual lexicon extraction from comparable corpora. We believe that our model is simple and sound. The most significant result is that the new approach to finding single word translations has been shown to be competitive and promising for future work. We hope that this new paradigm can lead to insights that could be unclear in other models. Dealing with this problem is an interesting line for future research.

Acknowledgments. The research leading to these results has received funding from the French National Research Agency under grant ANR-08-CORD-013.

References

1. Armstrong, S., Thompson, H.: A presentation of MLCC: Multilingual Corpora for Cooperation. In: Linguistic Database Workshop, Groningen (1995)
2. Chiao, Y.C., Zweigenbaum, P.: Looking for candidate translational equivalents in specialized, comparable corpora. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Tapei, Taiwan, pp. 1208–1212 (2002)
3. Chiao, Y.C., Zweigenbaum, P.: The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations. In: Baud, R., Fieschi, M., Le Beux, P., Ruch, P. (eds.) The New Navigators: from Professionals to Patients, Actes Medical Informatics Europe. Studies in Health Technology and Informatics, vol. 95, pp. 397–402. IOS Press, Amsterdam (2003)
4. Church, K.W., Mercer, R.L.: Introduction to the Special Issue on Computational Linguistics Using Large Corpora. Computational Linguistics 19(1), 1–24 (1993), `http://dblp.uni-trier.de`
5. Daille, B., Morin, E.: French-English Terminology Extraction from Comparable Corpora. In: Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP 2005), Jeju Island, Korea, pp. 707–718 (2005)
6. Déjean, H., Gaussier, E.: Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. Lexicometrica, Alignement lexical dans les corpus multilingues, pp. 1–22 (2002)
7. Déjean, H., Sadat, F., Gaussier, E.: An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Tapei, Taiwan, pp. 218–224 (2002)
8. Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics 19(1), 61–74 (1993)
9. Fano, R.M.: Transmission of Information: A Statistical Theory of Communications. MIT Press, Cambridge (1961)
10. Fung, P.: Compiling Bilingual Lexicon Entries From a non-Parallel English-Chinese Corpus. In: Farwell, D., Gerber, L., Hovy, E. (eds.) Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA 1995), Langhorne, PA, USA, pp. 1–16 (1995)
11. Fung, P.: A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In: Farwell, D., Gerber, L., Hovy, E. (eds.) AMTA 1998. LNCS (LNAI), vol. 1529, pp. 1–17. Springer, Heidelberg (1998)
12. Fung, P., Lo, Y.Y.: An ir approach for translating new words from nonparallel, comparable texts. In: Proceedings of the 17th International Conference on Computational Linguistics (COLING 1998), pp. 414–420 (1998)
13. Fung, P., McKeown, K.: Finding Terminology Translations from Non-parallel Corpora. In: Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC 1997), Hong Kong, pp. 192–202 (1997)
14. Gaussier, E., Renders, J.M., Matveeva, I., Goutte, C., Déjean, H.: A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona, Spain, pp. 526–533 (2004)

15. Gillard, L., Bellot, P., El-Bèze, M.: D'une compacité positionnelle à une compacité probabiliste pour un système de questions / réponses. In: CORIA, pp. 271–286 (2007)
16. Gillard, L., Sitbon, L., Blaudez, E., Bellot, P., El-Bèze, M.: Relevance Measures for Question Answering, The Lia at qa@clef-2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 440–449. Springer, Heidelberg (2007)
17. Grefenstette, G.: Corpus-Derived First, Second and Third-Order Word Affinities. In: Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX 1994), Amsterdam, The Netherlands, pp. 279–290 (1994)
18. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publisher, Boston (1994)
19. Hickl, A., Wang, P., Lehmann, J., Harabagiu, S.M.: Ferret: Interactive question-answering for real-world environments. In: ACL (2006)
20. Huang, Z., Thint, M., Qin, Z.: Question classification using head words and their hypernyms. In: EMNLP, pp. 927–936 (2008)
21. Laroche, A., Langlais, P.: Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, pp. 617–625 (2010)
22. Lavenus, K., Grivolla, J., Gillard, L., Bellot, P.: Question-answer matching: Two complementary methods. In: RIAO, pp. 244–259 (2004)
23. Morin, E., Daille, B., Takeuchi, K., Kageura, K.: Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic, pp. 664–671 (2007)
24. Peters, C., Picchi, E.: Cross-language information retrieval: A system for comparable corpus querying. In: Grefenstette, G. (ed.) Cross-Language Information Retrieval, ch.7, pp. 81–90. Kluwer Academic Publishers (1998)
25. Rapp, R.: Identify Word Translations in Non-Parallel Texts. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1995), Boston, MA, USA, pp. 320–322 (1995)
26. Rapp, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999), College Park, MD, USA, pp. 519–526 (1999)
27. Salton, G., Lesk, M.E.: Computer evaluation of indexing and text processing. Journal of the Association for Computational Machinery 15(1), 8–36 (1968)
28. Voorhees, E.M.: Overview of the trec 2004 question answering track. In: TREC (2004)