Word alignment applied on English-Arabic parallel corpus

Mourad Ellouze[1], Wafa Neifar[1,2], and Lamia Hadrich Belguith[1]

[1] ANLP-RG, MIRACL Laboratory, Sfax University, B.P-3018 Sfax, Tunisie
[2] LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
ellouzemourad@yahoo.fr, neifar@limsi.fr,lamia.belguith@gmail.com

Abstract. In this paper, we are interested to align simple and compound English-Arabic word from an English-Arabic medical corpus. Our goal is to improve the alignment results obtained by GIZA ++ tool. We thus propose a hybrid approach that uses, on the one hand, linguistic methods like morpho-syntactic tagging, syntactic patterns and transliteration detection, and statistical measures on the other hand such as mutual information, the harmonic mean, the likelihood coefficient and the χ^2. The evaluation of our system, we uses the Cambridge dictionary. The results obtained show that our proposed approach improves both the quality of the alignment and the translation.

Keywords: word alignment · parallel corpus · transliteration · Modern Standard Arabic

1 Introduction

The manual construction of bilingual and trilingual resources for low-resource languages such as Arabic is time-consuming and expensive. Through recent years, different methods and tools for the automatic processing of the Arabic language have been the subject of research thanks to the raising availability of parallel bilingual corpora. In this paper, we propose a hybrid approach to improve the quality of alignment produced by the GIZA ++ tool. Following a state of the art of different alignment methods in a general way in Section 2, we describe our approach implemented in Section 3. Then, we present and discuss the results obtained in Section 4 before concluding and providing some perspectives to this work in Section 5.

2 Related work

The problems of Bilingual text alignment at the word level are mainly related to the intrinsic characteristics of the language itself as well as to the different styles of writing. Also, many researches have led to the development of word alignment methods from parallel corpora. Other works have focused on the importance of treating simple and compound words like [12]. Their objective is to achieve a

hybrid approach to align simple and compound words and also idiomatic expressions from a parallel French-Arabic corpus. In the first place the method consists to perform morpho-syntactic labeling and lemmatization of the texts, then extract named Entities (EN) through a bilingual lexicon. The alignment of simple words is based on two alignments: i) first step of alignment on these ENs and words in the neighborhood using their corpus positions and grammatical labels attributed to each word, and ii) a second 'word alignment' step using the GIZA ++ tool [6]. The alignment of compound words uses a syntactic analysis of the corpus. The sequences of repetitive words are identified and the number of occurrences is calculated. This information is used to represent compound words in the form of vectors. A pairing is then performed using the "cos". A transliteration method has also been used to improve the alignment of proper names [2]. In order to take into account a complex and multifaceted situation, the approach consists of presenting the possible transliterations for each proper name from Arabic to French, using also a predefined dictionary. and the identification of cognates. The normalization of the generated words exploits the number of occurrences of the proper nouns returned by the Google search engine. [3] Propose a method for the alignment of complex terms extracted from each language from a parallel corpus of Italian-Arabic legal texts. For this, the index of words in the context is used as a translation relation indicator. A threshold applied to the distance between the translations makes it possible to filter the results.

3 Method of improving word level alignment

As indicated above, we propose a hybrid approach to improve the alignment of bilingual English-Arabic texts. We rely on a parallel English-Arabic corpus of the medical field [13]. This is a set of documents which are pamphlets of a few pages to the patients. Figure 1 shows the different stages of our work process.

3.1 Word Level Alignment

Alignment consists of matching the different linguistic units of the texts of two different languages with a translation relation. In our work, we used GIZA ++ [6] to achieve the initial alignment at the word level from a parallel English-Arabic corpus aligned at the sentence level. For several years, this tool has been recognized as the reference for performing word alignment. It's based on the models HMM probabilists to define alignment mappings. For each pair of parallel English-Arabic sentences, this step consists in associating for each word of the target language (Arabic language) the position of the set of words corresponding to the language source (English language). Figure 2 presents an extract of the result of this step. In order to be able to be exploited in the later stages, we have chosen to modify the display of the result of the alignment. Figure 3 presents the bilingual lexicon constructed from the source language word matching positions and the target language based on the previous results produced by GIZA ++.

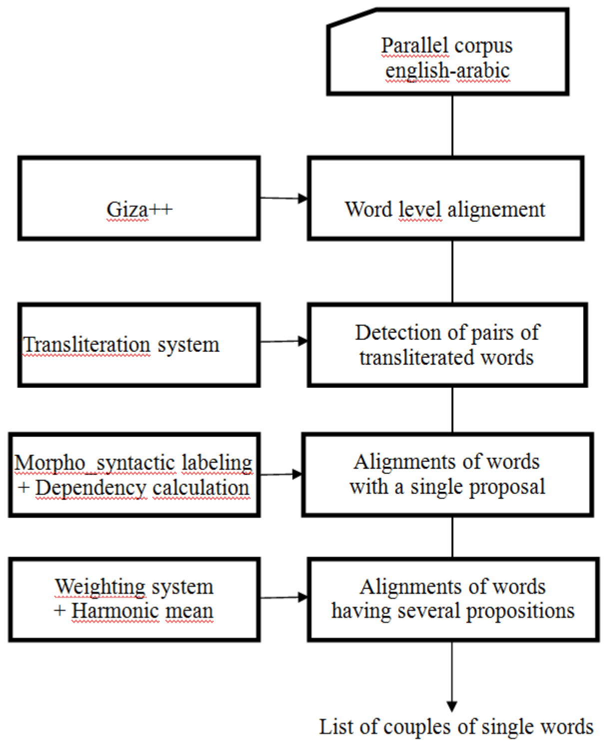

Fig. 1. Alignment process

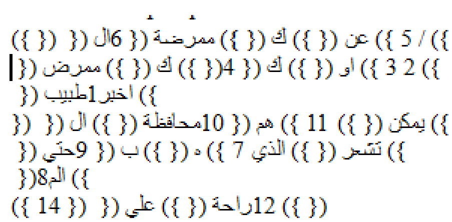

Fig. 2. Extract from the result of Giza ++

tell	اخبر
doctor	طبيب
your	ك
or	او
nurse	ممرض
about	عن
pain	الم
so	حتي
can	يمكن
they	هم
keep	محافظة
comfortable	راحة
you	ك
.	.

Fig. 3. Extract from our result of the alignment step

3.2 Detecting couples of transliterated words

The transliteration of words borrowed from Arabic consists in transcribing the words of a foreign language into Arabic characters. This strategy is often used for the proper names and more generally the named entities. So we used here the system produced by [11] which allows us to detect and extract pairs of medical terms transliterated from English into Arabic characters. In order for this system to be compatible with our objective and the results produced to respond our needs, we have applied some changes to its work process. Therefore , the modified transliteration system allows to identify the pairs composed from English words and their transliterated in Arabic. For now, it doesn't focus on medical terms. Figure 4 shows an excerpt from the transliteration step after adapting it to our work.

plastic	بلاستيك
routines	روتين
tofu	التوفو
cream	كريمات
testosterone	التيستوستيرون
kegel exercises	كيجل
kegel	كيجل
vitamin e	فيتامين
zoloft	زولوفت
paxil	باكسيل
neurontin	نيورونتين

Fig. 4. Extract of the transliteration result

3.3 Morpho-syntactic labeling

This step allows associating to each word of the text its corresponding grammatical category. If a word is aligned by GIZA ++ and is not validated by the transliteration step, we study morpho-syntactic labels and their grammatical features. We make the hypothesis that the translation of a noun is a noun. For this step, we used the Stanford POS Tagger system[5] for the English corpus and MADA + TOKEN[4] for the Arabic corpus. Figure 5 shows an extract from the labeling morpho-syntactic of English and Arabic corpora.

Fig. 5. Extract from the result of the morpho-syntactical labeling

3.4 Dependency calculation : Mesure of χ^2

Measure of Chi2 is a statistical measure which allows to calculate the degree of dependence between two words (see Equation 1 for the initial data presented in Table 1). If a word is not validated by the transliteration step and morpho-syntactic labeling, we calculate its Chi2. The null hypothesis is generally rejected when p <= 0.05.

$$\chi^2(i,j) = \frac{N(ad-cb)^2}{j_1 i_1 i_0 j_0} \tag{1}$$

Table 1. Initial data for calculating $\chi^2(i,j)$

	j	$\neg j$	
i	a	c	$i_1 = a + c$
$\neg i$	b	d	$i_0 = b + d$
	$j_1 = a + b$	$j_0 = c + d$	$N = a + b + c + d$

3.5 Weighting system

When several alignments are associated to word, we have applied the three most used measures: Mutual Information (Equation 2),χ^2(Equation 1) and Likelihood Coefficient (Equation 3). Mutual information [10] measures the statistical dependence of variables. This value is expressed by the ratio of the probability of observing i knowing that we observed j on the probability of i. The likelihood coefficient [7–9] is used in statistics and economics to compare two situations. So we have three weighted lists for each of the measures presented above.

$$IM(i,j) = log\frac{P(i|j)}{P(i)} \tag{2}$$

$$loglike(i,j) = \sum_{ij} \log \frac{k_{ij}N}{C_i R_j}$$

$$= k_{11} \log \frac{k_{11}N}{C_1 R_1} + k_{12} \log \frac{k_{12}N}{C_1 R_2} + k_{21} \log \frac{k_{21}N}{C_2 R_1} + k_{22} \log \frac{k_{22}N}{C_2 R_2} \tag{3}$$

where

k_{11} corresponds to the cooccurrences of the two words i and j
k_{12} is the difference between the number of occurrences of i and k_{11}
k_{21} is the difference between the number of occurrences of j and k_{11}
k_{22} is the total number of occurrences in the corpus - k_{12} - k_{21} + k_{11}

3.6 Reordering of candidates according to the harmonic mean

In order to combine the ranks of three weighted lists, we have reordered the simple candidate words according with the harmonic mean (see Equation 4).

$$MH(x,y,z) = \frac{3 \times x \times y \times z}{y \times z + x \times z + x \times y} \tag{4}$$

A post-processing allows us to eliminate the symbols and signs of punctuation as well as words of length equal to one letter.

3.7 Compound words

The alignment of simple words is insufficient because there are compound words or idiomatic expressions.
For example, *estrogen levels* translates literally and therefore corresponds to a word-for-word alignment. But, in French, the expression fige *pomme de terre* translates into a single word in Arabic. However, Giza ++ only offers word-for-word alignments or several words in the source language correspond to a single word in the target language. That is why we have proposed an approach to deal the cases of compound words where we start our work by validating the combinations of the source words judged as compound words by the Giza ++ tool in the first step we apply syntax patterns with the Nooj tool[14] and in

the second step we calculate the LLR value to measure the degree link between the different units of compound words. After having validated the source words composed by a linguistic and statistical method we pass to validate the couple which is composed by a compound source word and a simple target word because the result of Giza ++ does the alignment (N->1) for that we apply the Chi2 measure for each pair (English, Arabic). Then we proceed to treat the simple target words by looking for each unit of source word compound in our own bilingual lexicon (obtained from the alignment step of the simple words) if we find the correspondence of all the units in this case we translate them literally, otherwise we retain the proposition of Giza++ and we consider as an idiomatic expression. If in the case where we have a compound source word whose second unit is part of the list *up*, *down*, *out*, *off* we directly consider this word as an idiomatic expression that it must align with a single unit of target word. Finally, we decided to eliminate the sequences of five elements after an empirical study.

4 Experimental results

Our system has been tested on the Cambridge Dictionary. It is based on the Cambridge English Corpus, which contains more than 1.5 billion English words, and the Cambridge Learner Corpus. The performance of our system is evaluated by the three classical evaluation criteria : Precision, Recall and F-Measure.

Thus, we obtain a Precision of **0.59**, a Recall of **0.87** and an F-measure of **0.70** .We detail the results in Table 2.

Table 2. Validation of results

	Percentage
True positive	33,34%
False positive	22,96%
False negative	04,90%
True negative	38,77%

After the test phase of our corpus on GIZA ++, we obtained 6783 pairs of single words. After the intervention of our system, 2595 pairs are validated. The transliteration step validates 3.38% of the results. The step of calculating the Chi2 measure validates 53.71% and the morphological step validates 1.41%. The harmonic mean validates 41.48%. The final filtering step eliminated 3.53% of the results of our system. In table 3 we show the assesment of the result given by Giza ++ and our system.

Our system has validated 1330 pairs of compound words in which **73.38%** are correct.

Table 3. comparison of results for simple words.

	Giza++	Our system
Rate of accuracy	56.31%	**87.16%**
Error rate	43.68%	**12.83%**

5 Conclusion

In this paper, we have presented a hybrid approach of word alignment by combining statistical and linguistic information (transliteration, morphological analysis, and harmonic mean). The obtained results show that our approach makes it possible to improve the alignment produced by GIZA ++ from 56.31% to 87.16% of right answer. In future work, we plan to develop strategies and techniques which allows to update automatically the lexicon

References

1. Katerina T. Frantzi, and Sophia Ananiadou. (1999). The C–value/NC Value domain independent method for multi–word term extraction. In Journal of Natural Language Processing, 6(3), pp.145–179
2. Saadane Houda and Nasredine Semmar. (2012). Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe. In Actes de la conférence conjointe JEP- TALN-RECITAL, volume 2: TALN, Grenoble, pp.127–140.
3. Fawi Delmonte. 2015. (2015). Italian-Arabic domain terminology extraction from parallel corpora. In Proceedings of the Second Italian Conference on Computational Linguistics CliC-it 2015, pp.130 : Accademia University Press.
4. Nizar Habash. (2010). Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
5. Kristina Toutanova, Dan Klein, Christopher D. Manning and Yoram Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of HLT-NAACL 2003, pp.252–259.
6. Franz Josef Och and Hermann Ney (2003). A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1) pp.19–51, 2003.
7. Ted Dunning. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), pp.61–74.
8. Baayen, R. Harald. (2001). Word frequency distribution. Number 18 in Text, Speech and Language Technology. Dordrecht: Kluwer Academic Publishers.
9. Benoît Habert and Michèle Jardino. (2003). Compte rendu de R. Harald Baayen, Word Frequency Distribution. Traitement automatique des langues, 43(2), pp.209–211.
10. Claude E. Shannon. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, pp.379–423.
11. Wafa Neifar, Thierry Hamon, Pierre Zweigenbaum, Mariem Ellouze and Lamia Hadrich Belguith (2018). Détection des couples de termes translittérés à partir d'un corpus parallèle anglais-arabe. Actes de la conférence TALN 2018, pp.437–445.

12. Nasredine Semmar and Laib Meriama. (2012). Using a Hybrid Word Alignment Approach for Automatic Construction and Updating of Arabic to French Lexicons.Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pp.127–140.
13. Wafa Neifar, Thierry Hamon, Pierre Zweigenbaum, Mariem Ellouze Khemakhem, and Lamia Hadrich Belguith. (2016). Adaptation of a term extractor to Arabic specialised texts : First experiments and limits. In Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING2016), LNCS. Springer, April 2016.
14. Max Silberztein et Agnès Tutin(2005). NooJ, un outil TAL pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE. La revue Apprentissage des langues et systèmes d'information et de communication, Alsic. Vol. 8, n 2 — 2005 pp.123-134