العنوان:	A Study on Using English-Arabic Multiword Expressions for Statistical Machine Translation
المصدر:	مجلة التواصل اللساني
الناشر:	مؤسسة العرفان للإستشارات التربوية والتطوير المهني
المؤلف الرئيسي:	Bouamor, Dhouha
مؤلفين آخرين:	Zweigenbaum, Pierre، Semmar, Nasr Aldine(Advisor)
المجلد/العدد:	مج16, ملحق
محكمة:	نعم
التاريخ الميلادي:	2014
الصفحات:	7 - 20
رقم MD:	597158
نوع المحتوى:	بحوث ومقالات
اللغة:	English
قواعد المعلومات:	AraBase
مواضيع:	اللغة العربية، اللغة الإنجليزية، الترجمة الآلية
رابط:	http://search.mandumah.com/Record/597158

LINGUISTICA COMMUNICATIO: Vol.5, n°CS, 2013 ,p. 07-19

A study on Using English-Arabic MultiWord Expressions for Statistical Machine Translation

Dhouha Bouamor*, Nasredine Semmar* and Pierre Zweigenbaum**
***CEA, LIST, LVIC,**
F91191 Gif sur Yvette Cedex, France. Email: {prenom.nom@cea.fr}
**** LIMSI-CNRS, F-91403 Orsay, France. Email: pz@limsi.fr**

Abstract- Identifying and translating a Multi Word Expression (MWE) in a text represent an issue for numerous applications of Natural Language Processing (NLP) especially for Machine Translation (MT). In this paper, we describe an hybrid approach, combining linguistic and statistical information to extract and align MWEs from a sentence level aligned English -Arabic parallel corpus. In order to assess the quality of the mined bilingual MWEs, we conduct a Statistical Machine Translation (SMT) task-based evaluation. We investigate the performance of three methods aiming to integrate extracted bilingual MWEs in Moses, a phrase based SMT system. We experimentally show that these textual units enhance the translation quality for both In-Domain and Out-Of-Domain configurations.

Keywords- Mulit-Word Expressions (MWEs) extraction, Alignment of MWEs, Vector space model, Moses

Introduction

A Multi Word Expression (MWE) can be defined as a combination of words for which syntactic or semantic properties of the whole expression cannot be obtained from its parts [1], Such units are made up of collocations (ابتسامة عريضة big smile), compounds (الجلسة العامة plenary meeting), expressions more or less frozen as ضرب به عرض الحائط. which means in English ignore with contempt, named entities (البيت الأبيض, the White House) etc. [1], [2]. They are numerous and constitute a significant portion of the lexicon of any natural language. [3] claims that the frequency of MWEs in a speaker's lexicon is almost equivalent to the frequency of single words. While easily mastered by native speakers, their interpretation poses a major challenge for NLP applications especially those addressing semantic aspects of language.

For Statistical Machine Translation (SMT) systems, various improvements of translation quality were achieved with the emergence of phrase-based approaches [4]. Phrases are defined as simply arbitrary n-grams with no sophisticated linguistic motivation consistently translated in a parallel corpus. In such systems, the lack of an adequate processing of MWEs could affect he translation quality. In fact, the literal translation of an unrecognized expression by the system is the source of an erroneous and incomprehensible translation. For example, it would suggest city of amusement as a translation of مدينة الملاهي instead of amusement park. It is therefore important to make use a lexicon in which

MWEs are handled. But such kind of resource is not necessarily available in all languages, and if they exist, as described [5], they do not cover all MWEs of a given language.

In this paper, we consider any non-compositional contiguous sequence, belonging to one of the classes defined by [6], as a MWE. In [6], three classes of MWEs were distinguished on the basis of their categorical properties and their syntactic and semantic congealing degrees. They are made up of compounds, idiomatic expressions and collocations. Based on this classification, we present a method combining linguistic and statistical information to extract and align MWEs in an English-Arabic parallel corpus aligned at the sentence level. Intuitively, bilingual MWEs are useful to improve the performance of SMT. However, further research is still needed to find the best way to bring such external knowledge to the decoder. In this study, we view SMT as an extrinsic evaluation of the usefulness of MWEs and explore three different strategies for integrating such textual units in Moses, the state-of-the-art phrase based SMT system.

The remainder of this paper is organized as follows: the next section (section II) describe in some details previous works addressing the task of semantically equivalent translations extraction and their applications. In section III, we introduce a novel method for identifying Arabic and English MWEs and then present, in section IV, the algorithm we implemented to acquire translation pairs of MWEs and report our evaluation results. In section V three methods aiming to integrate MWEs in an SMT system are introduced and obtained results are discussed. We, finally, conclude and present our future work, in section VI.

Related Work

Numerous NLP approaches have already been introduced to deal with the problem of MWEs identification. Starting from a parallel corpus aligned at the sentence level, most works revolve around three approaches: (1) symbolic methods relying on morphosyntactic patterns; (2) statistical methods and (3) Hybrid approach combining (1) and (2). In [7], one of the earliest work using a symbolic approach, the author focused essentially on French-English noun groups identification. These textual units are recognized by exploiting their part-of-speech tag. Then, using an Expectation Maximization (EM) algorithm, bilingual correspondences are extracted. A precision value of 90% referred to the 100 first correspondences is reported. An extension of this method is proposed, later by [8] in order to detect MWEs by introducing a bidirectional version of the MWEs extraction algorithm. In this new version and to add prior information, the maximum likelihood estimate is replaced in the M-Step of the EM algorithm with the Maximum-A-Posteriori (MAP) estimate.

In another direction, [9] describe a semi-automatic tool, TERMIGHT, with the purpose of extracting technical noun groups using a syntactic pattern filter. They use a word alignment tool to align MWEs. For each source term, the tool identifies a candidate translation by selecting a sequence of target words whose first and last word are aligned with any of the words in the source term. The accuracy obtained for 192 English-German correspondences is about 40%.

In another perspective, some works attempt to extend the linguistic based methods, proposed for MWEs identification, by using, for example, additional association measures such as Mutual Information [10] or the Log Likelihood Ratio [11], [12] to capture the cohesion degree between the constituents of a MWE. However, these measures present two main shortcomings: they are designed only for bi-grams and trigrams and require a definition of a threshold above which an extracted phrase is considered as a MWE.

Furthermore, another type of heuristics, are applied for the MWEs alignment task. [13] and [12] claim that MWEs keep in most cases the same morphosyntactic structure in the source and target language, which is not universal. For example, the English MWE collective wisdom which is aligned with the Arabic MWE تحسين فعالية الوحدة does not share the same morphosyntactic structure.

The previous approaches are proposed to address the problem of MWEs extraction and alignment mainly in Latin languages (English, French, German,...). However, few works, focus on Arabic MWEs processing. For instance, [14] developed an hybrid multiword term extraction approach for Arabic in the specific domain of environment by combining grammatical patterns and statistical scores. More recently, [15] introduced three complementary approaches for the extraction of Arabic MWEs from different data sources: Arabic Wikipedia, English MWEs from Princeton WordNet 3.0 and a large unannotated corpus.

Most of the previous methods aim at identifying MWEs in a corpus to construct or extend a bilingual lexicon. Having such type of the textual units is useful for a variety of NLP applications such as information retrieval [16], word sense disambiguation [17]. A considerable amount of research has focused on the identification and extraction of MWEs in order to improve an MT system performance. [18] described an approach of noun-noun compound machine translation, but not significant comparison was presented. In [19], authors introduce a method in which a bilingual MWEs lexicon was used to modify the word alignment in order to improve the translation quality. In their work, bilingual MWEs were grouped as one unique token before training alignment models. They showed on a small corpus, that both alignment quality and translation accuracy were improved. However, in a further study, a lower BLEU score is reported after grouping MWEs by part-of-speech on a large corpus [20]. More recently, [21] described a method integrating an in-domain bilingual MWEs to Moses and gained +0.61 of BLEU score compared with the baseline system. In [22], [23], we proposed several methods to integrate a French-English bilingual lexicon of MWEs and report significant improvements in BLEU scores. In this paper, we apply the same techniques previously presented in [23] but focus on the English-Arabic language pair.

Monolingual extraction of MWEs

In this section, we describe the approach to extract monolingual MWEs from an English-Arabic parallel corpus. Generally, the choice of an effective way to deal with this problem depends on the further use of MWEs, and resources availability.

The method we present here is based on a symbolic approach relying on morphosyntactic patterns. Relatively simple, it handles with both frequent and infrequent expressions and do not use any dictionary. It only involves a full morphosyntactic analysis of source and target texts.

In the first step, we used the CEA LIST Multilingual Analysis platform LIMA [23] to analyze texts. After that, based on a list of morphosyntactic patterns, our method produces a list of monolingual Arabic and English MWEs.

Linguistic pre-processing

The LIMA linguistic analyzer is built using a traditional architecture involving separate processing modules:

- A Tokenizer which separates the input stream into a graph of words. This separation is achieved by an automaton developed for each language and a set of segmentation rules.
- A Morphological analyzer which looks up each word in a general full form dictionary. If these words are found, they are associated with their lemmas and all their grammatical tags. For Arabic agglutinated words which are not in the full form dictionary, a clitic stemmer was added to the morphological analyzer. The role of this stemmer is to split agglutinated words into proclitics, simple forms and enclitics.
- An Idiomatic Expressions recognizer which detects idiomatic expressions and considers them as single words for the rest of the processing. Idiomatic expressions are phrases or compound nouns that are listed in a specific dictionary. The detection of idiomatic expressions is performed by applying a set of rules that are triggered on specific words and tested on left and right contexts of the trigger. These rules can recognize contiguous expressions as (البيت الأبيض the White House in English). Noncontiguous expressions such as phrasal verbs are recognized too.
- A module to process unknown words by assigning to these words default linguistic properties based on features identified during tokenization (e.g. presence of Arabic or Latin characters, numbers, etc.).
- A Part-Of-Speech (POS) tagger which searches valid paths through all the possible tags paths using attested trigrams and bi-grams sequences. The trigram and bi-gram sequences are generated from a manually annotated training corpus. They are extracted from a hand-tagged corpora of 13 200 Arabic words. If no continuous trigram full path is found, the POS tagger tries to use bi-grams at the points where the trigrams were not found in the sequence. If no bi-grams allow completing the path, the word is left undisambiguated. The accuracy of the Arabic Part-Of-Speech tagger is around 91%.

Candidat identification

Since most MWEs consist of noun, adjectives, prepositions and determinant, we adopted for each language a linguistic filter keeping only n-gram units which match with a list of a hand created morphosyntactic patterns. Such process is used to keep only specific strings and filter out undesirable ones such as candidate composed mainly of stop words (of a, is a, that was). Our algorithm operates on lemmas instead of surface forms which can draw on richer statistics and overcome the data sparseness problems. In Table I we give an example of MWEs produced for each pattern.

We add to list of MWEs candidates, produced by our algorithm, the obtained list of named entities (Papua New Guinea, Korean peninsula etc.) recognized by the morphosyntactic analyzer. Then, we scored them using their total frequency in the corpus.

Arabic Patterns	Example
Noun-Noun	جزر سليمان
Noun-Det-Adj	جميع الأعمال
Noun-Det-Noun	حرية التعبير
Noun-Noun-Adj	إقامة سلام وطيد
N»mn-Adj-t)et-Noim	أضرار واسعة النطاق
Noun-Noun-Det-Noun	تعزر حقوق الطفل
Det-Noun-Dct-Adj	البيانات المتكره
Det-Noun-Det Non iv-Dei -Noun-Det-Noun	الدول الجزرية الصغرة النامية
English Pattern	**Example**
Adj-Noun	international law
Noun-Noun	working group
Past_PSrtidpleNouii	enforcod disappearance
Adj-Adj-Noun	flexible v»oiking arrangement
Adj-Noun-Adj	small island developing
Adj-Noun-Noun	international trade law
Noun-Prep-Noun	freedom of expression
Noun- Prcp-Noun-Noun	right of children identity
Adj-Noun-Prep-Noun	economic development ut' country

Table I

MORPHOSYNTACTIC CONFIGURATIONS FOR ARABIC AND ENGLISH MWES

To avoid an over-generation of MWEs and remove irrelevant candidates from the process, a redundancy cleaning approach is introduced. In this approach, if a MWE is nested in another, and they both have the same frequency, we discard the smaller one. Otherwise we keep both of them. We consider also the alternative of having a MWE that appear nested in a high number of terms. We followed [24] to discard all longer MWEs. It is important to note that, our approach does not use any additional correlation statistics such as Mutual Information or Log Likelihood Ratio since these measures require a definition of a threshold above of which an extracted phrase is considered as a MWE or not. Statistical methods have mostly been applied to bi-grams and trigrams and it becomes more problematic to extract MWEs of more than three words. Our method consider that all extracted units, regardless of their sizes, are effective and valid and includes all of them in the translation process. To our knowledge, it is the first time such n-grams are considered.

Vector space model for MWEs Alignment

We present a method aiming to find for each MWE in a source language its adequate translation in the target one. Traditionally, this task was handled through the use of external linguistic resources such as bilingual dictionaries or simple words alignment tools. We propose a resource-independant method which simply requires a parallel corpus and a list of input MWEs candidates to translate.

Our approach is based on aspects of the distributional semantics [25], where a specific representation is associated to each expression (source and target). We associate to each MWE an N sized vector, where N is the number of sentences in the corpus, indicating whether it appears or not in each sentence of the corpus. Our algorithm is based on the Vector Space Model (VSM). VSM [26] is a well-known algebraic model used in information retrieval, indexing and relevance ranking. This vector space representation will serve, eventually, as a basis to establish a translation relation between each pair of MWEs.

To extract translation pairs of MWEs, we propose an iterative alignment algorithm operating as follows:

1. Find the most frequent MWE exp in each source sentence.
2. Extract all target translation candidates, appearing in all parallel sentences to those containing exp.
3. Compute a confidence value VConf for each translation relation between exp and all target translation candidates.
4. Consider that the target MWE maximizing Vconf is the best translation.
5. Discard the translation pair from the process and go back to 1.

To compute the confidence value VConf, we adopted the Jaccard Index, a frequently used measure in information retrieval. It is defined as:

$$IJ = \frac{I_{st}}{V_s + V_t - I_{st}}$$

and based on the number I_{st} of sentences shared by each target and a source MWE. This is normalized by the sum of the number of sentences where the source and target MWEs appear independently of each other ($V_{s.}$ and V_t) decreased by I_a . Here is sample of aligned MWE by means of the algorithm described above.

freedom of thought →حرية الفكر

former Yugoslav → اليوغوسلافية السابقة

assistance mission→ بعثات تقدم المساعدة

developing world →العالم النامي

peacekeeping operations → عمليات حفظ السلام

illicit trade→التجارة غير المشروع

collective wisdom →تحسين فعالية الوحدة

ceasefire agreement →وقف أطلاق النار

Korean peninsula→ الجزيرة الكورية

By observing some pairs, we noticed that our method has two advantages: (1) It allows the translation of MWE aligned in most previous work [9], [21] using simple word alignment tools to establish word-to-word alignment relations. In our work, we capture the semantic equivalence between expressions such as developing world and العالم النامي in a different way. (2) It also permits the alignment of idioms such as peacekeeping operations → عمليات حفظ السلام.

We have also identified a class of error caused by the choice of n-gram's size. Since our system does not capture one-to-many correspondences, some MWEs were not aligned correctly.

Application of MWEs

In the previous section, we described the approach we followed to extract translation pairs of MWEs. Because of the lack of a common benchmark data sets for evaluation in MWE extraction and alignment research, we decided to study in what respect these units are useful to improve the performance of phrase based SMT systems. In such systems, phrase tables are the main knowledge source for the machine translation decoder.

The decoder consults these tables to figure out how to translate an input candidate in a source language in the target one. However, due to the errors in automatic word alignment, extracted phrases could be meaningless. To alleviate this problem, we propose three techniques to make use of bilingual MWEs in an SMT system and compare their performances.

Methods

Retraining model with MWEs

In this method (noted BASELINE +T RAIN), we add the extracted bilingual MWE as a parallel corpus and retrain the model. By increasing the occurrences of bilingual MWEs, considered as good phrases, we expect a modification of alignment and probability estimation.

MWEs in the phrase table

Here we attempt to extend an SMT system's phrase table by integrating the found bilingual MWEs candidates [(1)] . We, then use the Jaccard Index (proposed for each pairs of MWE) to define the two directions translation probabilities and set the lexical probabilities to 1 for simplicity. So, for each phrase in a given input sentence, the decoder will search all candidate translation phrases by taking into account bilingual MWEs. This method is denoted BASELINE-STABLE in the remaining part of this paper.

New feature to MWEs

[27] pointed out that better feature mining can lead to substantial gain in translation quality. We followed this claim and extend BASELINE-STABLE by adding a new feature indicating whether a phrase is a MWE or not. The aim of this method (BASELINE+FEAT) is to guide the system to choose bilingual MWEs instead of its phrases.

Data

In order to constitute our training corpus, we extracted 50000 sentence pairs from the Resolution of the Unite Nations General Assembly [28]. This corpus regroups a set of English-Arabic parallel sentences belonging to six principal organs: the general Assembly, the Security Council, the Economic and Social Council, the Trusteeship Council, the International Court of Justice, and the Secretariat. Table II describe Training and Test corpus characteristics.

Data	Arabic	English
Training-set	50000	
Words	105762	103531
In-Domain Test	1000	
Words	24565	22763
Out-Of-Domain Test	1000	
Words	27546	23975

Table II: Characteristics of training and test data

We conducted two test experiments: In-Domain and Out-Of-Domain. For this, we randomly extracted 1000 parallel sentences from the corpus described above as an In-

([1])The MWEs extracted following the approach w present in section IV

Domain corpus and 1000 pairs of sentences from news. This type of study is generally done to show the impact of the domain vocabulary on the translation results.

First, training and test corpora were tokenized. For Arabic, we used the Toolkit AMIRA [(2)], a suite of tools for the processing of Modem Standard Arabic texts. It takes the Buckwalter transliteration input encoding formats and produces segmented output. For English, the corpus was tokenized using OpenNLP [(3)]. Then, we cleaned up the training corpus and only kept sentences containing at most 50 words. We used the tokenized Arabic sentences in the training set to construct a five-gram language model. This model was trained by employing the IRST Language Modeling Toolkit [(4)].

We also extracted MWEs from this training corpus and applied the three methods described above. We, consequently, exploited the full list of available resources.

Baseline

We use Moses [4], an open source SMT system, as our baseline system {BASELINE). When dealing with Arabic, most works consider only the Arabic to English translation direction. In this work, we present experiments done in the other direction: English to Arabic Moses system. For this, we make use it as a phrase based translation model in which a translation table contain both single words and phrases.

The features used in baseline system include :(1) four translation probability features, (2) one language model and (3) word penalty. For the BASELJNE+TRAIN method, bilingual MWEs are added into the training corpus, as result, new alignment and phrase table are obtained. For BASELINE-STABLE method, bilingual units are incorporated in the Baseline system's phrase table. For BASELINE+FEAT method, an additional 1/0 feature is introduced to each entry of the phrase table.

Results and discussion

We measure translation quality on the two test sets described in the previous section and calculate the BLEU score. We also consider only one reference for each test sentence. Obtained results are reported in Table III.

Methods	**BLEU_Score**	
	In-Domain	**Out-Of-Domain**
BASELINE	56.13	2.14
BASELINE+TRAIN	56.33	2.22
BASELINE+TABLE	56.12	2.17
BASELINE+FEAT	54.92	2.39

Table III: Translation Results in term of BLEU score

(2) http://www.ibridgenetwork.org/columbia/ir_ms-12s-2-1

(3)http://opennlp.apache.org/

(4) http://hlt.fbk.eu/en/irstlm

The first substantial observation is that, when the test set is In-Domain, we achieve a relatively high score BLEU for all methods. For instance, in the BASELINE +TRAIN method exploiting only MWEs as additional parallel "sentences", we report an improvement of +0.20 points in BLEU score. Compared to the BASELINE system, the BASELINE+TABLE method have a slightly lower BLEU score. This may be due to the cost of using "short sentences", MWEs in this case. In fact, using longer sentences is less costly in the translation process even if the propose sentence is semantically invalid.

For the Out-Of-Domain test corpus, it is not surprising that our methods perform worst than in the In-Domain test set with a very low score BLEU. This result can be explained by the fact of using a different corpus with different vocabulary in the train which is different from the lexicon used in the test set. The most important result is that bilingual MWEs improve translation quality in all cases. The best improvement of the BLEU score is achieved using BASELINE+FEAT with a gain of +0.25 compared to BASELINE. This result shows the impact of adding the feature guiding the SMT system in choosing the best translation with a preference to the MWEs. It also, give us an idea about the role of such kind of lexical units integration in improving an SMT system performance. The BASELINE+TRAIN comes next with +0.08 BLEU score improvement. Finally, when using BASELINE+TABLE, we report a gain of +0.03 BLEU score.

The question that arises based on these different results is: Is it possible to claim that the system having the best score is the best one? In other words, are the obtained results for the different experimental settings statistically significant?

In order to assess statistical significance of previously obtained test results, we use the paired bootstrap resampling method. This method estimates the probability (p-value) that a measured difference in BLEU scores arose by chance by repeatedly (10 times) creating new virtual test sets by drawing sentences with replacement from a given collection of translated sentences. If there is no significant difference between the systems (i.e., the null hypothesis is true), then this shuffling should not change the computed metric score. We carry out experiments using this method to compare each of the methods BASELINE+TRAIN, BASELINE+TABLE and BASELINE+FEAT, yielding improvements in BLEU scores (Table III) over the BASELINE system on the two test set results In-Domain and Out-Of-Domain.

Methods	**p-value 95%CI**	
	In-Domain	**Out-Of-Domain**
BASELINE	-	-
BASELINE+TRAIN	0,3	0,2
BASELINE+TABLE	-	0,01
BASELINE+FEAT	-	0,01

Table IV: Statistical significance tests of BLEU improvements in term of p-value.

Table displays reported p-values at the edge of the 95\% confidence interval (Cl). As can be observed, the results vary from insignificant (at p>0.05) to highly significant. On

the Out-Of-Domain test set, we notice that improvements achieved by the BASELINE+FEAT and BASELINE+TABLE integration strategies are statistically significant. However, the small improvements of BLEU scores yielded by the BASELINE+TRAIN method having a p-value of 0.3 and 0.2 on respectively the In-Domain and Out-Of-Domain test sets are not significant.

Conclusion

We described, in this paper a method aiming to extract and align MWEs in an English-Arabic parallel corpus. The alignment algorithm we propose works only on many to many correspondences and deal with both frequent and infrequent MWEs in a given sentence pair.

We also investigated the performance of three different application strategies by integrating bilingual MWEs in the Moses SMT system. Results show that when the test set belongs to the same domain on which the translation model was trained, using MWEs as additional semantic information does not improve the translation quality and even if a small improvement is yielded, the later is not significant (i.e. improvement reported by the BASELINE+TRAIN method).

However, when the test corpus is in another domain, it is important to add some additional features to significantly improve the SMT system's performance.

Although our initial experiments are positive, we believe that they can be improved in a number of ways. We first plan to extend the morphosyntactic patterns to handle with other forms of MWEs, e.g. starting with a verb. We will also try to develop and evaluate other statistical based methods to align MWEs. In addition to their application in a phrase based SMT system, bilingual MWEs may also be integrated into other MT models such as rule-based translation ones. We also expect to extract such textual units from more available but less parallel data sources: comparable corpora.

Acknowledgment

This research work is supported by FINANCIAL WATCH (QNRF NPRP: 08-583-1-101) project. This publication was made possible by a grant from the Qatar National Research Fund NPRP 08-583-1-101. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the QNRF.

References

I. Sag, T. Baldwin, F. Francis Bond, A. Copestake, and D. Flickinger, “Multiword expressions:a pain in the neck for NLP,” in CICLing 2002, Mexico City, Mexico, 2002.

M. Constant, I. Tel Her, D. Duchier, Y. Dupont, A. Sigogne, S. Billot et al., “Integrer des connaissances linguistiques dans un CRF: application a l’apprentissage d’un segmenteur- etiqueteur du fran^ais,” in Actes de TALN, Montpellier, France, 2011.

R. Jackendoff, “The architecture of the language faculty,” MIT Press, 1997.

P. Koehn, F. Och, and D. Marcu, “Statistical phrase-based translation,” in Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, Canada, 2003, pp. 115-124.

B. Sagot, L. Clément, E. De La Clergerie, P. Boullier et al., “Vers un méta-lexique pour le franipais: architecture, acquisition, utilisation,” in Actes de TALN, 2005.

L. Nerima, V. Seretan, and E. Wehrli, “Le probelme de collocation en Tal,” in Nouveaux cahiers de linguistiques FranÇaise, 2006, pp. 95-115

J. Kupiec, “An algorithm for finding noun phrases correspondences in bilingual corpora.” in Proceedings of the 31st annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, USA, 1993, pp. 17-22.

T. Okita, M. Guerra, Y. Alfredo Graham, and A. Way, “Multi-word expression sensitive word alignment,” in Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010, Beijing, 2010, pp. 26-34.

I. Dagan and K. Church, “Termight: Identifying and translating technical terminology,” in Proceedings of the 4th Conference on ANLP, Stuttgart, Germany, 1994, pp. 34-40.

B. Daille, “Extraction de collocation partir de textes,” in Actes de TALN 2001 (Traitement automatique des langues naturelles), D. Maurel, Ed., ATALA. Tours: University de Tours, Jul. 2001.

C. Wu and S. J. Chang, “Bilingual collocation extraction based on syntactic and statistical analyses,” in Computational Linguistics, 2004, pp. 1-20.

V. Seretan and E. Wehrli, “Collocation translation based on sentence alignment and parsing,” in Actes de TALN 2007 (Traitement automatique des langues naturelles), F. Benarmara, N. Hatout, P. Muller, and S. Ozdowska, Eds., ATALA. Toulouse: IRIT, Jun. 2007.

I. Tufis and R. Ion, “Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure,” in Proceedings of the 4th International Conference on Speech and Dialogue Systems, 2007, pp. 183-195

S. Boulaknadel, B. Daille, and A. Driss, “A multi-term extraction program for arabic language,” in Proceedings of LREC, Marrakech, Morocco, 2008.

M. Attia, A. Toral, L. Tounsi, P. Pecina, and J. Van Genabith, “Automatic extraction of multiword expressions,” in Proceedings of the Workshop on MultiWord Expressions: From theory to application., Beiging, 2010.

O. Vechtomova, "The role of multi-word units in interactive information retrieval," in ECIR2005, Berlin, 2005, pp. 403-420.

M. Finlayson and N. Kulkami, "Detecting multi-word expressions improves word sense disambiguation," in Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, Portland, Oregon, USA, 2011, pp. 20-24.

T. Tanaka and T. Baldwin, "Noun-noun compound machine translation: A feasibility study on shallow processing," in Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, 2003.

P. Lambert and R. Banchs, "Data inferred multi-word expressions for statistical machine translation," in Proceedings of MT SUMMIT, 2005.

P. Lambert and R. Banchs, "Grouping multi-word expressions according to part-of-speech in statistical machine translation," in Proceedings of the Workshop on Multi-word Expressions in a multilingual context, 2006.

Z. Ren, Y. Lu, Q. Liu, and Y. Huang, "Improving statistical machine translation using domain bilingual multiword expressions," in Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications, 2009, pp. 47-57.

D. Bouamor, N. Semmar, and P. Zweigenbaum, "Improved statistical machine translation using multi-word expressions," in Proceedings of MT-LIHMT, Barcelona, Spain, 2011.

D. Bouamor, N. Seminar, and P. Zweigenbaum, "Identifying bilingual multi-word expressions for statistical machine translation," in Proceedings of LREC, Istanbul, turkey, 2012.

R. Besanon, G. De Chalendar, O. Ferret, F. Gara, M. Laib, O. Mesnard, and N. Semmar, "Lima: A multilingual framework for linguistic analysis and linguistic resources development and evaluation," in Proceedings of LREC, Malta, 2010.

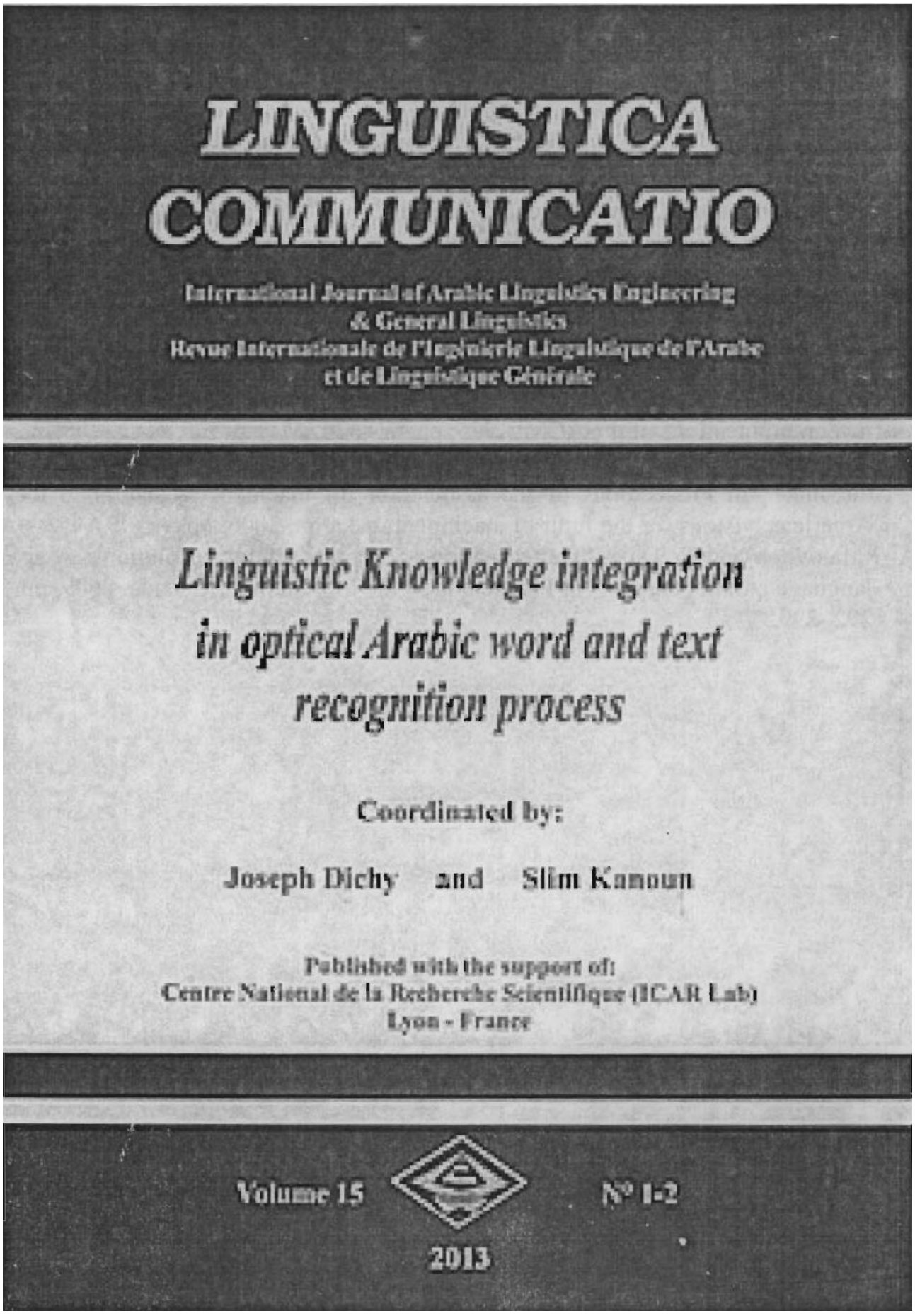
C. Frantzie, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the c-value/nc-value method," in Int. J. on Digital Libraries 3(2), 2000, pp. 115-130.

Z. Harris, "Distributional structure." Word, 1954.

G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," in Communications of the ACM, 1975, pp. 61-620.

A. Lopez and P. Resnik, "Word-based alignment, phrase based translation: what's the link?" in Proceedings of the association for machine translation in the Americas: visions for the futur of machine translation, 2006, pp. 90-99.

A. Rafalovitch and R. Dale, "United nations general assembly resolutions: A six language parallel corpus," in Proceeding of the MT-Summit, Canada, 2009, pp. 292-299.


LINGUISTICA COMMUNICATIO
International Journal of Arabic Linguistics Engineering & General Linguistics
Revue Internationale de l'Ingénierie Linguistique de l'Arabe et de Linguistique Générale
Linguistic Knowledge integration in optical Arabic word and text recognition process
Coordinated by:
Joseph Dichy and Slim Kanoun
Published with the support of:
Centre National de la Recherche Scientifique (ICAR Lab)
Lyon - France
Volume 15
N° 1-2
2013