Extraction of Multi-word Expressions from Small Parallel Corpora

Yulia Tsvetkov
Department of Computer Science
University of Haifa
yulia.tsvetkov@gmail.com

Shuly Wintner
Department of Computer Science
University of Haifa
shuly@cs.haifa.ac.il

Abstract

We present a general methodology for extracting multi-word expressions (of various types), along with their translations, from small parallel corpora. We automatically align the parallel corpus and focus on *mis*alignments; these typically indicate expressions in the source language that are translated to the target in a non-compositional way. We then use a large monolingual corpus to rank and filter the results. Evaluation of the quality of the extraction algorithm reveals significant improvements over naïve alignment-based methods. External evaluation shows an improvement in the performance of machine translation that uses the extracted dictionary.

1 Introduction

Multi-word Expressions (MWEs) are lexical items that consist of multiple orthographic words (e.g., *ad hoc*, *by and large*, *New York*, *kick the bucket*). MWEs are numerous and constitute a significant portion of the lexicon of any natural language. They are a heterogeneous class of constructions with diverse sets of characteristics, distinguished by their idiosyncratic behavior. Morphologically, some MWEs allow some of their constituents to freely inflect while restricting (or preventing) the inflection of other constituents. In some cases MWEs may allow constituents to undergo non-standard morphological inflections that they would not undergo in isolation. Syntactically, some MWEs behave like words while other are phrases; some occur in one rigid pattern (and a fixed order), while others permit various syntactic transformations. Semantically, the compositionality of MWEs is gradual, ranging from fully compositional to idiomatic (Bannard et al., 2003).

Because of their prevalence and irregularity, MWEs must be stored in lexicons of natural language processing applications. Handling MWEs correctly is beneficial for a variety of applications, including information retrieval, building ontologies, text alignment, and machine translation.

Identifying MWEs and extracting them from corpora is therefore both important and difficult. In Hebrew (which is the subject of our research), this is even more challenging due to two reasons: the rich and complex morphology of the language; and the dearth of existing language resources, in particular parallel corpora, semantic dictionaries and syntactic parsers.

We propose a novel algorithm for identifying MWEs in bilingual corpora, using automatic word alignment as our main source of information. In contrast to existing approaches, we do not limit the search to one-to-many alignments, and propose an error-mining strategy to detect misalignments in the parallel corpus. We also consult a large monolingual corpus to rank and filter out the expressions. The result is fully automatic extraction of MWEs of various types, lengths and syntactic patterns, along with their translations. We demonstrate the utility of the methodology on Hebrew-English MWEs by incorporating the extracted dictionary into an existing machine translation system.

The main contribution of the paper is thus a new alignment-based algorithm for MWE extraction that focuses on misalignments, augmented by validating statistics computed from a monolingual corpus. After discussing related work, we detail in Section 3 the methodology we propose. Section 4 provides a thorough evaluation of the results. We then extract translations of the identified MWEs and evaluate the contribution of the extracted dictionary in Section 5. We conclude with suggestions for future research.

Coling 2008: Poster Volume, pages 1256–1264,
Beijing, August 2010

2 Related Work

Early approaches to identifying MWEs concentrated on their collocational behavior (Church and Hanks, 1989). Pecina (2008) compares 55 different association measures in ranking German Adj-N and PP-Verb collocation candidates. This work shows that combining different collocation measures using standard statistical classification methods improves over using a single collocation measure. Other results (Chang et al., 2002; Villavicencio et al., 2007) suggest that some collocation measures (especially PMI and Log-likelihood) are superior to others for identifying MWEs. Soon, however, it became clear that mere co-occurrence measurements are not enough to identify MWEs, and their linguistic properties should be exploited as well (Piao et al., 2005). Hybrid methods that combine word statistics with linguistic information exploit morphological, syntactic and semantic idiosyncrasies to extract idiomatic MWEs.

Semantic properties of MWEs can be used to distinguish between compositional and non-compositional (idiomatic) expressions. Katz and Giesbrecht (2006) and Baldwin et al. (2003) use Latent Semantic Analysis for this purpose. They show that compositional MWEs appear in contexts more similar to their constituents than non-compositional MWEs. Van de Cruys and Villada Moirón (2007) use unsupervised learning methods to identify non-compositional MWEs by measuring to what extent their constituents can be substituted by semantically related terms. Such techniques typically require lexical semantic resources that are unavailable for Hebrew.

An alternative approach to using semantics capitalizes on the observation that an expression whose meaning is non-compositional tends to be translated into a foreign language in a way that does not result from a combination of the literal translations of its component words. Alignment-based techniques explore to what extent word alignment in parallel corpora can be used to distinguish between idiomatic expressions and more transparent ones. A significant added value of such works is that MWEs can thus be both identified in the source language and associated with their translations in the target language.

Villada Moirón and Tiedemann (2006) focus on Dutch expressions and their English, Spanish and German translations in the Europarl corpus (Koehn, 2005). To extract the candidates, they use syntactic properties (based on full parsing of the Dutch text) and statistical association measures. This approach requires syntactic resources that are unavailable for Hebrew.

Some recent works concentrate on exploiting translational correspondences of MWEs from (small) parallel corpora. MWE candidates and their translations are extracted as a by-product of automatic word alignment of parallel texts. Unlike Villada Moirón and Tiedemann (2006), who use aligned parallel texts to *rank* MWE candidates, Caseli et al. (2009) actually use them to extract the candidates. After the texts are word-aligned, Caseli et al. (2009) extract sequences of length 2 or more in the source language that are aligned with sequences of length 1 or more in the target. Candidates are then filtered out of this set if they comply with pre-defined part-of-speech patterns, or if they are not sufficiently frequent in the parallel corpus. Even with the most aggressive filtering, precision is below 40% and recall is extremely low (F-score is below 10 for all experiments). Our setup is similar, but we extract MWE candidates from the aligned corpus in a very different way; and we use statistics collected from a *monolingual* corpus to filter and rank the results.

Zarrieß and Kuhn (2009) also use aligned parallel corpora but only focus on one-to-many word alignments. To restrict the set of candidates, they focus on specific syntactic patterns as determined by parsing both sides of the corpus (again, using resources unavailable to us). The results show high precision but very low recall.

3 Methodology

We propose an alternative approach to existing alignment-based techniques for MWE extraction. Using a small bilingual corpus, we extract MWE candidates from noisy word alignments in a novel way. We then use statistics from a large monolingual corpus to rank and filter the list of candidates. Finally, we extract the translation of candidate MWEs from the parallel corpus and use them in a machine translation (MT) system.

3.1 Motivation

Parallel texts are an obvious resource from which to extract MWEs. By definition, idiomatic expressions have a non-compositional meaning, and hence may be translated to a single word (or to an expression with a different meaning) in a foreign language. The underlying assumption of alignment-based approaches to MWE extraction is that MWEs are aligned across languages in a way that differs from compositional expressions; we share this assumption. However, existing approaches focus on the results of word alignment in their quest for MWEs, and in particular consider 1:n and n:m alignments as potential areas in which to look for MWEs. This is problematic for two reasons: first, word alignment algorithms have difficulties aligning MWEs, and hence 1:n and n:m alignments are often noisy; while these environments provide cues for identifying MWEs, they also include much noise. Second, our experimental scenario is such that our parallel corpus is particularly small, and we cannot fully rely on the quality of word alignments, but we have a bilingual dictionary that compensates for this limitation. In contrast to existing approaches, then, we focus on *misalignments*: we trust the quality of 1:1 alignments, which we verify with the dictionary; and we search for MWEs exactly in the areas that word alignment *failed* to properly align, not relying on the alignment in these cases.

Moreover, in contrast to existing alignment-based approaches, we also make use of a large monolingual corpus from which statistics on the distribution of word sequences in Hebrew are drawn. This has several benefits: of course, monolingual corpora are easier to obtain than parallel ones, and hence tend to be larger and provide more accurate statistics. Furthermore, this provides validation of the MWE candidates that are extracted from the parallel corpus: rare expressions that are erroneously produced by the alignment-based technique can thus be eliminated on account of their low frequency in the monolingual corpus.

Specifically, we use pointwise mutual information (PMI) as our association measure. While PMI has been proposed as a good measure for identifying MWEs, it is also known not to discriminate accurately between MWEs and other frequent collocations. This is because it promotes collocations whose constituents rarely occur in isolation (e.g., typos and grammar errors), and expressions consisting of some word that is very frequently followed by another (e.g., *say that*). However, such cases do not have idiomatic meanings, and hence at least one of their constituents is likely to have a 1:1 alignment in the parallel corpus; we only use PMI *after* such alignments have been removed.

An added value of our methodology is the automatic production of an MWE translation dictionary. Since we start with a parallel corpus, we can go back to that corpus after MWEs have been identified, and extract their translations from the parallel sentences in which they occur.

Finally, alignment-based approaches can be symmetric, and ours indeed is. While our main motivation is to extract MWEs in Hebrew, a byproduct of our system is the extraction of *English* MWEs, along with their translations to Hebrew. This, again, contributes to the task of enriching our existing bilingual dictionary.

3.2 Resources

Our methodology is in principle language-independent and appropriate for medium-density languages (Varga et al., 2005). We assume the following resources: a small bilingual, sentence-aligned parallel corpus; large monolingual corpora in both languages; morphological processors (analyzers and disambiguation modules) for the two languages; and a bilingual dictionary. Our experimental setup is Hebrew-English. We use a small parallel corpus (Tsvetkov and Wintner, 2010) consisting of 19,626 sentences, mostly from newspapers. The corpus consists of 271,787 English tokens (14,142 types) and 280,508 Hebrew tokens (12,555 types), and is similar in size to that used by Caseli et al. (2009).

We also use data extracted from two monolingual corpora. For Hebrew, we use the morphologically-analyzed MILA corpus (Itai and Wintner, 2008) with part-of-speech tags produced by Bar-Haim et al. (2005). This corpus is much larger, consisting of 46,239,285 tokens (188,572 types). For English we use Google's Web 1T corpus (Brants and Franz, 2006).

Finally, we use a bilingual dictionary consisting of 78,313 translation pairs. Some of the entries were collected manually, while others are produced automatically (Itai and Wintner, 2008; Kirschenbaum and Wintner, 2010).

3.3 Preprocessing the corpora

Automatic word alignment algorithms are noisy, and given a small parallel corpus such as ours, data sparsity is a serious problem. To minimize the parameter space for the alignment algorithm, we attempt to reduce language specific differences by pre-processing the parallel corpus. The importance of this phase should not be underestimated, especially for alignment of two radically different languages such as English and Hebrew (Dejean et al., 2003).

Hebrew,[1] like other Semitic languages, has a rich, complex and highly productive morphology. Information pertaining to gender, number, definiteness, person, and tense is reflected morphologically on base forms of words. In addition, prepositions, conjunctions, articles, possessives, etc., may be concatenated to word forms as prefixes or suffixes. This results in a very large number of possible forms per lexeme. We therefore tokenize the parallel corpus and then remove punctuation. We analyze the Hebrew corpus morphologically and select the most appropriate analysis in context. Adopting this selection, the surface form of each word is reduced to its base form, and bound morphemes (prefixes and suffixes) are split to generate stand-alone "words". We also tokenize and lemmatize the English side of the corpus, using the Natural Language Toolkit package (Bird et al., 2009).

Then, we remove some language-specific differences automatically. We remove frequent function words: in English, the articles a, an and the, the infinitival to and the copulas am, is and are; in Hebrew, the accusative marker at. These forms do not have direct counterparts in the other language.

For consistency, we pre-process the monolingual corpora in the same way. We then compute the frequencies of all word bi-grams occurring in each of the monolingual corpora.

3.4 Identifying MWE candidates

The motivation for our MWE identification algorithm is the assumption that there may be three sources to misalignments (anything that is not a 1:1 word alignment) in parallel texts: either MWEs (which trigger 1:n or n:m alignments); or language-specific differences (e.g., the source language lexically realizes notions that are realized morphologically, syntactically or in some other way in the target language); or noise (e.g., poor translations, low-quality sentence alignment, and inherent limitations of word alignment algorithms).

This motivation induces the following algorithm. Given a parallel, sentence-aligned corpus, it is first pre-processed as described above, to reduce the effect of language-specific differences. We then use Giza++ (Och and Ney, 2003) to word-align the text, employing *union* to merge the alignments in both directions. We look up all 1:1 alignments in the dictionary. If the pair exists in our bilingual dictionary, we remove it from the sentence and replace it with a special symbol, '*'. Such word pairs are not parts of MWEs. If the pair is not in the dictionary, but its alignment score is very high (above 0.5) and it is sufficiently frequent (more than 5 occurrences), we add the pair to the dictionary but also retain it in the sentence. Such pairs are still candidates for being (parts of) MWEs.

Example 1 *Figure 1-a depicts a Hebrew sentence with its word-by-word gloss, and its English translation in the parallel corpus. Here,* bn adm *"person" is a MWE that cannot be translated literally. After pre-processing (Section 3.3), the English is represented as "and i tell her keep away from person" (note that* to *and* the *were deleted). The Hebrew, which is aggressively segmented, is represented as in Figure 1-b. Note how this reduces the level of (morphological and orthographic) difference between the two languages. Consequently, Giza++ finds the alignment depicted in Figure 1-c. Once 1:1 alignments are replaced by '*', the alignment of Figure 1-d is obtained.*

If our resources were perfect, i.e., if word alignment made no errors, the dictionary had perfect coverage and our corpora induced perfect statis-

[1] To facilitate readability we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are *abgdhwzxTiklmns'pcqršt*.

a.	*wamrti*	*lh*	*lhzhr*	*mbn*	*adm*	*kzh*
	and-I-told	to-her	to-be-careful	from-child	man	like-this

“and I told her to keep away from the person”

b.	*w*	*ani*	*amr*	*lh*	*lhzhr*	*m*	*bn*	*adm*	*k*	*zh*
	and	I	tell	to-her	to-be-careful	from	child	man	like	this

c.	*w*	*ani*	*amr*	*lh*	*lhzhr*	*m*	*bn adm*	*k*	*zh*
	and	I	told	her	keep away	from	person	{}	{}

d.	*	*	*	*	*lhzhr*	*	*bn adm*	*k*	*zh*
	*	*	*	*	keep away	*	person		

Figure 1: Example sentence pair (a); after pre-processing (b); after word alignment (c); and after 1:1 alignments are replaced by ‘*’ (d)

tics, then all remaining text (other than the special symbol) in the parallel text would be part of MWEs. In other words, all sequences of remaining source words, separated by ‘*’, are MWE candidates. As our resources are far from perfect, further processing is required in order to prune these candidates. For this, we use association measures computed from the monolingual corpus.

3.5 Ranking and filtering MWE candidates

The algorithm described above produces sequences of Hebrew word forms (free and bound morphemes produced by the pre-processing stage) that are not 1:1-aligned, separated by ‘*’s. Each such sequence is a MWE candidate. In order to rank the candidates we use statistics from a large *monolingual* corpus. We do *not* rely on the alignments produced by Giza++ in this stage.

We extract all word bi-grams from the remaining candidates. Each bi-gram is associated with its PMI-based score,[2] computed from the monolingual corpus. Interestingly, about 20,000 candidate MWEs are removed in this stage because they do not occur at all in the monolingual corpus.

We then experimentally determine a threshold (see Section 4). A word sequence *of any length* is considered MWE if all the adjacent bi-grams it contains score above the threshold. Finally, we restore the original forms of the Hebrew words in the candidates, combining together bound morphemes that were split during pre-processing; and we restore the function words. Many of the candidate MWEs produced in the previous stage are eliminated now, since they are not genuinely multi-word in the original form.

Example 2 *Refer back to Figure 1-d. The sequence bn adm k zh is a MWE candidate. Two bi-grams in this sequence score above the threshold: bn adm, which is indeed a MWE, and k zh, which is converted to the original form kzh and is hence not considered a candidate. We also consider adm k, whose score is low. Note that the same aligned sentence can be used to induce the* English *MWE keep away, which is aligned to a single Hebrew word.*

3.6 Results

As an example of the results obtained with this setup, we list in Table 1 the 15 top-ranking extracted MWEs. For each instance we list an indication of the type of MWE: person name (PN), geographical term (GT), noun-noun compound (NNC) or noun-adjective combination (N-ADJ). Of the top 100 candidates, 99 are clearly MWEs,[3] including *mzg awir* (*temper-of air*) “weather”, *kmw kn* (*like thus*) “furthermore”, *bit spr* (*house-of book*) “school”, *šdh t‘wph* (*field-of flying*) “airport”, *tšwmt lb* (*input-of heart*) “attention”, *ai apšr* (*not possible*) “impossible” and *b‘l ph*

[2] PMI^k is a heuristic variant of the PMI measure, proposed and studied by Daille (1994), where k, the exponent, is a frequency-related factor, used to demote collocations with low-frequency constituents. The value of the parameter k can be chosen freely ($k > 0$) in order to tune the properties of the PMI to the needs of specific applications. We conducted experiments with k = 0, 0.1, ... , 3 and found k = 2.7 to give the best results for our application.

[3] This was determined by two annotators.

(*in-on mouth*) “orally”. Longer MWEs include *ba lidi biTwi* (*came to-the-hands-of expression*) “was expressed”; *xzr ‘l ‘cmw* (*returned on itself*) “recurred”; *ixd ‘m zat* (*together with it*) “in addition”; and *h‘crt hkllit šl haw”m* (*the general assembly of the UN*) “the UN general assembly”.

Hebrew	Gloss	Type
xbr hknst	MP	NNC
tl abib	Tel Aviv	GT
gwš qTip	Gush Katif	NNC-GT
awpir pins	Ophir Pines	PN
hc‘t xwq	Legislation	NNC
axmd Tibi	Ahmad Tibi	PN
zhwh glawn	Zehava Galon	PN
raš hmmšlh	Prime Minister	NNC
abšlwm wiln	Avshalom Vilan	PN
br awn	Bar On	PN
mair šTrit	Meir Shitrit	PN
limwr libnt	Limor Livnat	PN
hiw‘c hmšpTi	Attorney General	N-ADJ
twdh rbh	thanks a lot	N-ADJ
rcw‘t ‘zh	Gaza Strip	NNC-GT

Table 1: Results: extracted MWEs

4 Evaluation

MWEs are notoriously hard to define, and no clear-cut criteria exist to distinguish between MWEs and other frequent collocations. In order to evaluate the utility of our methodology, we conducted three different types of evaluations that we detail below and in Section 5.

First, we use a small annotated corpus of Hebrew noun-noun constructions that was made available to us (Al-Haj and Wintner, 2010). The corpus consists of 463 high-frequency bi-grams of the same syntactic construction; of those, 202 are tagged as MWEs (in this case, noun compounds) and 258 as non-MWEs. This corpus consolidates the annotation of three annotators: only instances on which all three agreed were included. Since it includes both positive and negative instances, this corpus facilitates a robust evaluation of precision and recall. Of the 202 positive examples, only 121 occur in our parallel corpus; of the 258 negative examples, 91 occur in our corpus. We therefore limit the discussion to those 212 examples whose MWE status we can determine, and ignore other results produced by the algorithm we evaluate.

On this corpus, we compare the performance of our algorithm to four baselines: using only PMI to rank the bi-grams in the parallel corpus; using PMI computed from the monolingual corpus to rank the bi-grams in the parallel corpus; and using Giza++ 1:n alignments, ranked by their PMI (with bi-grams statistics computed once from parallel and once from monolingual corpora). ‘MWE’ refers to our algorithm. For each of the above methods, we set the threshold at various points, and count the number of true MWEs above the threshold (true positives) and the number of non-MWEs above the threshold (false positives), as well as the number of MWEs and non-MWEs below the threshold (false positives and true negatives, respectively). From these four figures we compute precision, recall and their harmonic mean, f-score, which we plot against (the number of results above) the threshold in Figure 2. Clearly, the performance of our algorithm is consistently above the baselines.

Second, we evaluate the algorithm on additional datasets. We compiled three small corpora of Hebrew two-word MWEs. The first corpus, **PN**, contains 785 person names (names of Knesset members and journalists), of which 157 occur in the parallel corpus. The second, **Phrases**, consists of 571 entries beginning with the letter *x* from a dictionary of Hebrew phrases (Rosenthal, 2009), and a set of 331 idioms we collected from internet resources. Of those, 154 occur in the corpus. The third set, **NN**, consists of the positive examples in the annotated corpus of noun-noun constructions described above.

Since we do not have negative examples for these sets, we only evaluate recall, using a threshold reflecting 2750 results. For each of these datasets, we report the number of MWEs in the dataset (which also occur in the parallel corpus, of course) our algorithm detected. We compare in Table 2 the recall of our method (MWE) to Giza++ alignments, as above, and list also the upper bound (UB), obtained by taking all above-threshold bi-grams in the corpus.

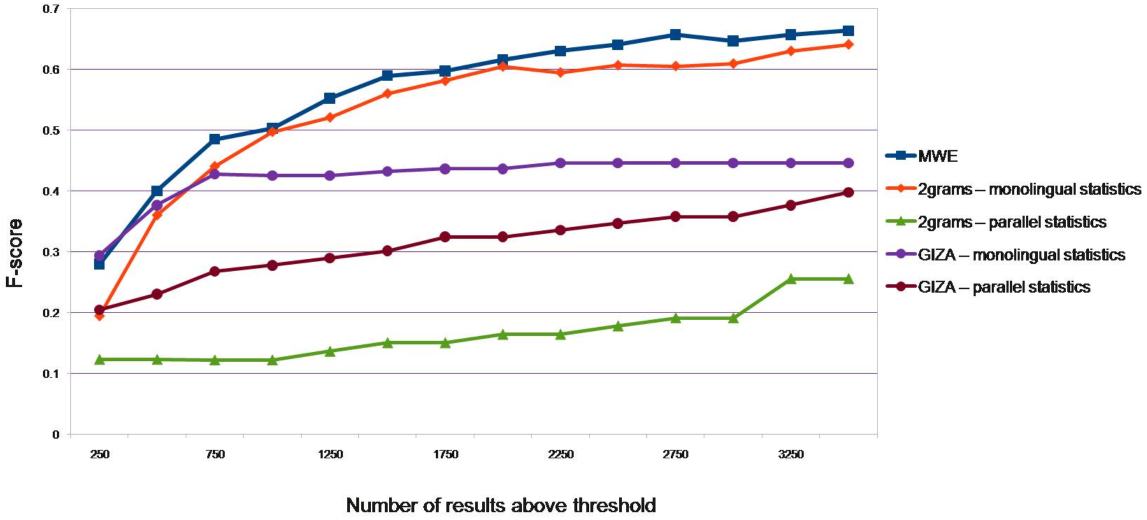

Figure 2: Evaluation results compared with baselines: noun-noun compounds

Method	**PN**		**Phrases**		**NN**	
	#	%	#	%	#	%
UB	74	100	40	100	89	100
MWE	66	89.2	35	87.5	67	75.3
Giza	7	9.5	33	82.5	37	41.6

Table 2: Recall evaluation

5 Extraction of MWE translations

An obvious benefit of using parallel corpora for MWE extraction is that the translations of extracted MWEs are available in the corpus. We use a naïve approach to identify these translations. For each MWE in the source-language sentence, we consider as translation all the words in the target-language sentence (in their original order) that are aligned to the word constituents of the MWE, as long as they form a contiguous string. Since the quality of word alignment, especially in the case of MWEs, is rather low, we remove "translations" that are longer than four words (these are most often wrong). We then associate each extracted MWE in Hebrew with all its possible English translations.

The result is a bilingual dictionary containing 2,955 MWE translation pairs, and also 355 translation pairs produced by taking high-quality 1:1 word alignments (Section 3.4). We used the extracted MWE bilingual dictionary to augment the existing (78,313-entry) dictionary of a transfer-based Hebrew-to-English statistical machine translation system (Lavie et al., 2004b). We report in Table 3 the results of evaluating the performance of the MT system with its original dictionary and with the augmented dictionary. The results show a statistically-significant ($p < 0.1$) improvement in terms of both BLEU (Papineni et al., 2002) and Meteor (Lavie et al., 2004a) scores.

Dictionary	BLEU	Meteor
Original	13.69	33.38
Augmented	13.79	33.99

Table 3: External evaluation

As examples of improved translations, a sentence that was originally translated as "His teachers also hate to the Zionism and besmirch his HRCL and Gurion" (fully capitalized words indicate lexical omissions that are transliterated by the MT system) is translated with the new dictionary as "His teachers also hate to the Zionism and besmirch his Herzl and David Ben-Gurion"; a phrase originally translated as "when so" is now properly translated as "likewise"; and several occurrences of "down spring" and "height of spring" are corrected to "Tel Aviv".

6 Conclusion

We described a methodology for extracting multiword expressions from parallel corpora. The algorithm we propose capitalizes on semantic cues provided by ignoring 1:1 word alignments, and viewing all other material in the parallel sentence as potential MWE. It also emphasizes the importance of properly handling the morphology and orthography of the languages involved, reducing wherever possible the differences between them in order to improve the quality of the alignment. We use statistics computed from a large monolingual corpus to rank and filter the results. We used the algorithm to extract MWEs from a small Hebrew-English corpus, demonstrating the ability of the methodology to accurately extract MWEs of various lengths and syntactic patterns. We also demonstrated that the extracted MWE bilingual dictionary can improve the quality of MT.

This work can be extended in various ways. While several works address the choice of association measure for MWE identification and for distinguishing between MWEs and other frequent collocations, it is not clear which measure would perform best in our unique scenario, where candidates are produced by word (mis)alignment. We intend to explore some of the measures discussed by Pecina (2008) in this context. The algorithm used for extracting the translations of candidate MWEs is obviously naïve, and we intend to explore more sophisticated algorithms for improved performance. Also, as our methodology is completely language-symmetric, it can be used to produce MWE candidates in English. In fact, we already have such a list of candidates, whose quality we will evaluate in the future. Finally, as our main motivation is high-precision, high-recall extraction of Hebrew MWEs, we develop other, non-alignment-based approaches to the task (Al-Haj and Wintner, 2010), and would like to explore the utility of combining different approaches to the same task under a unified framework. We are actively pursuing these research directions.

Acknowledgments

This research was supported by THE ISRAEL SCIENCE FOUNDATION (grants No. 137/06, 1269/07). We are grateful to Hassan Al-Haj for providing the noun compound annotated corpus and to Gennadi Lembersky for his help with the machine translation system.

References

Al-Haj, Hassan and Shuly Wintner. 2010. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, August.

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 89–96. Association for Computational Linguistics.

Bannard, Colin, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In Francis Bond, Anna Korhonen, Diana McCarthy and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72.

Bar-Haim, Roy, Khalil Sima'an, and Yoad Winter. 2005. Choosing an optimal architecture for segmentation and POS-tagging of Modern Hebrew. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 39–46, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA.

Brants, Thorsten and Alex Franz. 2006. Web 1T 5-gram version 1.1. LDC Catalog No. LDC2006T13.

Caseli, Helena, Aline Villavicencio, André Machado, and Maria José Finatto. 2009. Statistically-driven alignment-based multiword expression identification for technical domains. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 1–8, Singapore, August. Association for Computational Linguistics.

Chang, Baobao, Pernilla Danielsson, and Wolfgang Teubert. 2002. Extraction of translation unit from Chinese-English parallel corpora. In *Proceedings of the first SIGHAN workshop on Chinese language processing*, pages 1–5, Morristown, NJ, USA. Association for Computational Linguistics.

Church, Kenneth. W. and Patrick Hanks. 1989. Word association norms, mutual information and lexicography (rev). *Computational Linguistics*, 19(1):22–29.

Daille, Béatrice. 1994. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Université Paris 7.

Dejean, Herve, Eric Gaussier, Cyril Goutte, and Kenji Yamada. 2003. Reducing parameter space for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts*, pages 23–26, Morristown, NJ, USA. Association for Computational Linguistics.

Itai, Alon and Shuly Wintner. 2008. Language resources for Hebrew. *Language Resources and Evaluation*, 42:75–98, March.

Katz, Graham and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia, July. Association for Computational Linguistics.

Kirschenbaum, Amit and Shuly Wintner. 2010. A general method for creating a bilingual transliteration dictionary. In *Proceedings of The seventh international conference on Language Resources and Evaluation (LREC-2010)*, May.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X, Phuket, Thailand.*

Lavie, Alon, Kenji Sagae, and Shyamsundar Jayaraman. 2004a. The significance of recall in automatic metrics for mt evaluation. In Frederking, Robert E. and Kathryn Taylor, editors, *AMTA*, volume 3265 of *Lecture Notes in Computer Science*, pages 134–143. Springer.

Lavie, Alon, Shuly Wintner, Yaniv Eytani, Erik Peterson, and Katharina Probst. 2004b. Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In *Proceedings of TMI-2004: The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Pecina, Pavel. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*.

Piao, Scott Songlin, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for mwe extraction. *Computer Speech and Language*, 19(4):378–397.

Rosenthal, Ruvik. 2009. *Milon HaTserufim (Dictionary of Hebrew Idioms and Phrases)*. Keter, Jerusalem. In Hebrew.

Tsvetkov, Yulia and Shuly Wintner. 2010. Automatic acquisition of parallel corpora from websites with dynamic content. In *Proceedings of The seventh international conference on Language Resources and Evaluation (LREC-2010)*, May.

Van de Cruys, Tim and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic, June. Association for Computational Linguistics.

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP'2005*, pages 590–596.

Villada Moirón, Begoña and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context*. Association for Computational Linguistics.

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043.

Zarrieß, Sina and Jonas Kuhn. 2009. Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 23–30, Singapore, August. Association for Computational Linguistics.