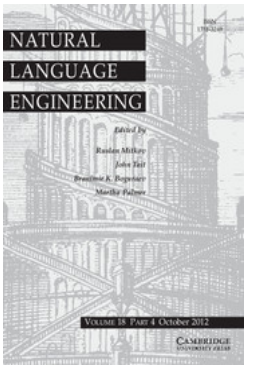

Extraction of multi-word expressions from small parallel corpora

YULIA TSVETKOV and SHULY WINTNER

Natural Language Engineering / Volume 18 / Issue 04 / October 2012, pp 549 - 573
DOI: 10.1017/S1351324912000101, Published online:

Natural Language Engineering **18** (4): 549–573. © Cambridge University Press 2012
doi:10.1017/S1351324912000101

Extraction of multi-word expressions from small parallel corpora

YULIA TSVETKOV[1] and SHULY WINTNER[2]
[1]*Language Technologies Institute Carnegie Mellon University, Pittsburgh, PA, USA*
e-mail: yulia.tsvetkov@gmail.com
[2]*Department of Computer Science University of Haifa, Hafia, Israel*
e-mail: shuly@cs.haifa.ac.il

(*Received 28 March 2011; revised 22 February 2012; accepted 22 February 2012; first published online 21 March 2012*)

Abstract

We present a general, novel methodology for extracting multi-word expressions (MWEs) of various types, along with their translations, from small, word-aligned parallel corpora. Unlike existing approaches, we focus on *mis*alignments; these typically indicate expressions in the source language that are translated to the target in a non-compositional way. We introduce a simple algorithm that proposes MWE candidates based on such misalignments, relying on 1:1 alignments as anchors that delimit the search space. We use a large monolingual corpus to rank and filter these candidates. Evaluation of the quality of the extraction algorithm reveals significant improvements over naïve alignment-based methods. The extracted MWEs, with their translations, are used in the training of a statistical machine translation system, showing a small but significant improvement in its performance.

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 a heterogeneous class of constructions with diverse sets of characteristics, distinguished by their idiosyncratic behavior (see Section 2). Morphologically, some MWEs allow some of their constituents to freely inflect while restricting (or even 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 others 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 fully idiomatic (Bannard, Baldwin and Lascarides 2003).

Multi-word Expressions are extremely prevalent: the number of MWEs in a speaker's lexicon is estimated to be of the same order of magnitude as the number of single words (Jackendoff 1997). This may even be an underestimate, as 41% of the entries in WordNet 1.7 (Fellbaum 1998), for example, are multi-words (Sag *et al.* 2002). An empirical study (Erman and Warren 2000) found that over 55% of the

tokens in the studied texts were instances of *prefabs* (defined informally as word sequences that are preferred by native speakers because of conventionalization).

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 (Doucet and Ahonen-Myka 2004), building ontologies (Venkatsubramanyan and Perez-Carballo 2004), text alignment (Venkatapathy and Joshi 2006), and machine translation (MT) (Baldwin and Tanaka 2004; Uchiyama, Baldwin and Ishizaki 2005). Identifying MWEs and extracting them from corpora is therefore both important and difficult. In this work we focus on Hebrew,[1] in which this task 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 unsupervised algorithm for identifying MWEs in (small) 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 only address continuous MWEs in this work, whose meaning is non-compositional; but they can be of varying lengths.) We demonstrate the utility of the methodology on Hebrew–English MWEs by incorporating the extracted dictionary into an existing MT system.

This paper is a revised and extended version of Tsvetkov and Wintner (2010b), adding a much more detailed discussion of the task and of related work, a better presentation of the methodology, several additional experiments that establish the robustness of our results and the individual contribution of some of the sub-stages of our methodology, and a detailed error analysis.

We discuss some properties of Hebrew MWEs in Section 2, and describe related work in Section 3. Section 4 details the main methodology and results, and a robust evaluation is provided in Section 5. We conclude with suggestions for future research.

2 Hebrew MWEs

Multi-word Expressions exhibit several properties that make them both interesting and challenging for processing, and Hebrew is no different in this respect. In this section we briefly recapitulate some of those properties, focusing on syntax and semantics, and exemplify them on Hebrew, following Al-Haj (2010). We also define the task we address in this work by constraining the types of MWEs that our solution identifies.

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

2.1 Properties

First, MWEs occur in various syntactic constructions:

Noun-noun *bit xwlim*
house-of patients
"patient house" ⟹ hospital

Noun-adjective *sprwt iph*
literature pretty
"beautiful literature" ⟹ belles-lettres

Adjective-noun *išr lb*
straight-of heart
"straight hearted" ⟹ honest

Participle-noun *'wrk din*
editor-of law
"law editor" ⟹ lawyer

Verb-preposition *mt 'l*
die on
"die on" ⟹ be in love with

Conjunction *ala am kn*
but if thus
"but if so" ⟹ unless

Proper name *brq awbmh*
Barack Obama
"Barack Obama" ⟹ Barack Obama

Second, some MWEs are fixed lexical combinations, while others are more flexible. As an example of the former consider

ap 'l pi š
even on mouth–of that
"even on the mouth that" ⟹ although

In this idiomatic expression, the constituents and the order in which they occur in a text are fixed. In contrast, the following expression

išb 'l X šb'h
sat on X seven
"sit seven (days) on someone" ⟹ mourn

contains an open slot that can be filled by a noun phrase, and the order of the two objects can be changed.

Semantically, MWEs cover a wide spectrum, from highly idiomatic

iwca dwpn
go-out side,bank
"leaving through the membrane" ⟹ exceptional

to completely transparent

'wbd zr
worker foreign
"foreign worker" $\Longrightarrow$ foreign worker

2.2 Task definition

As shown above, MWEs are a diverse set of constructions, exhibiting a variety of linguistic phenomena, with various idiosyncratic properties. Our task in this work is to identify such constructions in textual corpora. However, we do not attempt to extract all of them.

First, we only address *continuous* MWEs. That is, our solution will not identify expressions with "slots" in them that can be filled by productive phrases (such as *išb 'l X šb'h* "sat on X seven"). Second, our solution will likely fail to identify MWEs whose meaning is compositional. Our methodology capitalizes on the non-compositionality of many MWEs, and examples such as *'wbd zr* "worker foreign" are unlikely to be extracted.

Our solution is not limited to any combination of part-of-speech categories. Specifically, we will identify several of the constructions listed above, including proper names. While many consider named entity recognition to be a separate problem, we maintain that proper names are a special kind of MWEs (which may or may not be easier to identify than other kinds). Note that Hebrew does not use capitalization, which makes the recognition of named entities harder than in European languages. In addition, many proper names are derived from (and are homonymous with) common nouns, again adding to the difficulty of identifying them. For example, *bt im* (*daughter-of sea*) "mermaid" is also the name of a city in Israel (*Bat Yam*); both are spelled the same, i.e., with no capitalization.

Finally, our task is to identify MWEs of *any* length. While much of our processing is done for bi-grams (sequences of two tokens), the methodology works equally well for longer sequences, as our results in Section 4.6, and in particular Table 4, demonstrate. Note that many of the results include function words, even though we remove (very few, extremely frequent) function words as part of our pre-processing (Section 4.3).

3 Related work

Early approaches to MWE identification concentrated on their collocational behavior (Church and Hanks 1990). One of the first approaches was implemented as Xtract (Smadja 1993): Here, word pairs that occur with high frequency within a context of five words in a corpus are first collected, and are then ranked and filtered according to contextual considerations, including the parts of speech of their neighbors.

Pecina (2008) compares fifty-five 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, Danielsson and Teubert 2002; Villavicencio *et al.* 2007) suggest that some collocation measures (especially PMI and Log-likelihood) are in fact superior to others for identifying

MWEs. Co-occurrence measures alone are probably not enough to identify MWEs, and the linguistic properties of such expressions should be considered as well (Piao *et al.* 2005). Hybrid methods that combine word statistics with linguistic information exploit morphological, syntactic, and semantic idiosyncratic properties of MWEs to identify them in corpora.

Cook, Fazly, and Stevenson (2007), for example, use prior knowledge about the overall syntactic behavior of an idiomatic expression to determine whether an instance of the expression is used literally or idiomatically. They assume that in most cases, idiomatic usages of an expression tend to occur in a small number of canonical forms for that idiom; in contrast, the literal usages of an expression are less syntactically restricted, and are expressed in a greater variety of patterns, involving inflected forms of the constituents.

Al-Haj and Wintner (2010) focus on morphological idiosyncrasies of Hebrew MWEs, and leverage such properties to automatically identify a specific construction, noun–noun compounds, in a given text. However, they do not rely on the semantics of MWEs, which is the focus of our current research. Moreover, the approach cannot be easily extended to cover other syntactic constructions, whereas our method is construction-independent.

Semantic properties of MWEs can be used to distinguish between compositional and non-compositional (idiomatic) expressions. Baldwin *et al.* (2003) and Katz and Giesbrecht (2006) use Latent Semantic Analysis (LSA) for this purpose. They show that compositional MWEs appear in contexts more similar to their constituents than non-compositional MWEs. For example, the co-occurrence measured by LSA between the expression *'kick the bucket'* and the word *die* is much higher than co-occurrence of this expression and its component words. The disadvantage of this methodology is that to distinguish between idiomatic and non-idiomatic usages of the MWE it relies on the MWE's known idiomatic meaning, and this information is usually absent. In addition, this approach fails when only idiomatic or only literal usages of the MWE are overwhelmingly frequent.

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. MWE candidates and their translations are extracted as a by-product of automatic word alignment of parallel texts (Och and Ney 2003).

Lambert and Banchs (2005) define phrases that are hard to align as *bilingual multi-word expressions*. They use an asymmetry-based approach and focus on alignment

sets in which source-to-target links proposed by Giza++ (Och and Ney 2003) are different from target-to-source alignments. They then amend word alignments according to the alignment mismatches they detect and show that translation quality improves. Whereas the motivation of Lambert and Banchs (2005) is to improve MT, ours is to extract MWEs; consequently, they evaluate their method only in the context of an MT system, whereas we provide both intrinsic and extrinsic evaluations. Finally, our method can work with a relatively small parallel corpus, compensating with a larger monolingual corpus.

Villada Moirón and Tiedemann (2006) focus on Dutch expressions and their English, Spanish, and German translations in the Europarl corpus (Koehn 2005). MWE candidates are ranked by the variability of their constituents' translations. To extract the candidates, they use syntactic properties (based on full parsing of the Dutch text) and statistical association measures. Translational entropy (Melamed 1997) is used as the main criterion for distinguishing between idiomatic expressions and non-idiomatic ones. This approach requires syntactic resources that are unavailable for Hebrew.

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 (*m*:*n* alignments). Candidates are then filtered out of this set if they comply with predefined 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: We do not assume that sequences of m words in the source language are necessarily aligned with n words in the target. Rather, all we require is that these sequences *not* be 1:1 aligned in order for them to be considered candidates (in particular, we also consider words aligned to *null*). We consult a dictionary to validate 1:1 alignments; 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.

Ren *et al.* (2009) extract MWEs from the source side of a parallel corpus, ranking candidates on the basis of a collocation measure (log-likelihood). They then word-align the parallel corpus and naïvely extract the translations of candidate MWEs based on the results of the aligner. To filter out the list of translations, they use a classifier informed by "translation features" and "language features" (roughly corresponding to the translation models and language models used in MT). The extracted translation pairs are fed into a baseline Chinese–English MT system and improve BLEU results by up to 0.61 points. While our MWE extraction algorithm is very different, and our translation extraction method is more naïve,

we, too, use MT as an extrinsic evaluation method for the quality of the extracted translations.

More recently, Carpuat and Diab (2010) proposed two different strategies for integrating MWEs in MT systems: A static integration strategy that segments training and test sentences according to the MWE vocabulary; and a dynamic integration strategy that adds a new MWE-based feature to the phrase table used by MT systems. This dynamic feature represents the number of MWEs in the input language phrase, and is a generalization of the binary MWE feature of Ren *et al.* (2009). The evaluation shows that these two strategies are complementary, and both of them improve English–Arabic translation quality. Similarly, we show that our rather naïve integration of an MWE dictionary in an MT system improves its performance.

4 Extracting MWEs from parallel corpora

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 an MT system.

4.1 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 (some, typically more idiomatic) MWEs are aligned across languages in a way that differs from other, 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 them. 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 (for example, they can consist of fragments of MWEs, sometimes with additional unrelated material). 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. In other words, we view all words that are not included in 1:1 alignments as potential areas in which to search for MWEs, independently of how these words were aligned

by the word-aligner. In particular, we also consider words that are aligned to *null* in such contexts. Unlike other alignment-based approaches, then, our algorithm is less susceptible to noise, first because we validate 1:1 alignments with a dictionary, and second, because our focus on misalignments improves the chances of aligning chunks that *include* multi-word expressions, rather than smaller chunks that may consist of proper substrings thereof.

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 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 a variant of 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 our approach is indeed symmetric. While our main motivation is to extract MWEs in Hebrew, a by-product 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.

4.2 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 2010a), which consists of 19,626 sentences, mostly from newspapers. Some data on the parallel corpus are listed in Table 1 (the size of our corpus is very similar to that of 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, Sima'an and Winter (2005). For English we use

Table 1. *Statistics of the parallel corpus*

	English	Hebrew
Number of tokens	271,787	280,508
Number of types	14,142	12,555
Number of unique bi-grams	132,458	149,668

Table 2. *Statistics of the Hebrew corpus*

Number of tokens	46,239,285
Number of types	188,572
Number of unique bi-grams	5,698,581

Google's Web 1T corpus (Brants and Franz 2006). Data on the Hebrew corpus are provided in Table 2.[2]

Finally, we use a bilingual dictionary consisting of 78,313 translation pairs. Most of the entries were collected manually (Itai and Wintner 2008), while few were produced automatically from Wikipedia article titles (Kirschenbaum and Wintner 2010).

4.3 Pre-processing 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). See also Section 5.2.

Hebrew, 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. Consequently, a single English word (e.g., the noun *advice*) can be aligned to hundreds or even thousands of Hebrew forms (e.g., *l'cth* "to-her-advice"). As *advice* occurs only eight times in our small parallel corpus, it would be almost impossible to collect statistics even on simple 1:1 alignments without appropriate tokenization and lemmatization.

We therefore tokenize the parallel corpus and then remove punctuation. We analyze the Hebrew corpus morphologically and use a disambiguation module to

[2] Web-scale data such as the Google Web 1T corpus are unavailable for Hebrew. While web-extracted counts were shown to be informative in the absence of large monolingual corpora (Lapata and Keller 2005; Nakov and Hearst 2005), our Hebrew corpus is sufficiently large, so we had no need to resort to harvesting noisy data from the web.

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, Klein and Loper 2009). Then, we try to 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 either do not have direct counterparts in the other language, or behave very differently across the languages.

Example 1
Following is an example Hebrew sentence from our corpus with a word-by-word gloss and an English translation:

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"

After pre-processing, the Hebrew sentence, which is aggressively segmented, is represented as follows:

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

The English sentence is represented as *and i tell her keep away from person* (note that *to* and *the* are deleted). Note how this reduces the level of (morphological and orthographic) difference between the two languages.

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.

4.4 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., one language lexically realizes notions that are realized morphologically, syntactically, or in some other way in the other 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 as produced by Giza++ is very high (above 0.5) and it is sufficiently frequent (more than five 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.[3]

Example 2
Refer back to Example 1. Following is the representation of the two sentences after pre-processing, and the alignment produced by Giza++. Sequences that are aligned to a single word in the other language are enclosed in curly brackets; and *null* alignments are indicated by {}.

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

Once 1:1 alignments are replaced by '*', the following alignment is obtained

*	*	*	*	*lhzhr*	*	{*bn adm*}	*k*	*zh*
*	*	*	*	{keep away}	*	person		

Note that we are not concerned about the actual alignments of remaining tokens; unlike other approaches, that focus only on *m*:*n* alignments, we generalize to other cases of misalignments, including those in which words in one language are aligned to *null*. Specifically, in the example above, the bigram *bn adm* is considered a MWE candidate independently of the English words its tokens are aligned with.

If our resources were perfect, i.e., if word alignment made no errors, the dictionary had perfect coverage and our corpora induced perfect statistics, then all the remaining text (other than the special symbol) in the parallel text would be part of MWEs. In other words, all sequences of remaining source-language 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.

4.5 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 *contiguous* sequence of tokens, unbroken by '*'s, 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 these candidates (contiguous token sequences). Each bi-gram is associated with its PMI-based score, computed from the monolingual corpus. We use PMI^k, a heuristic variant of the PMI measure, proposed and studied by Daille (1994). The exponent, k, 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

[3] The thresholds were determined without empirical experimentation. We believe that fine-tuning of these parameters, maximizing the accuracy on a development corpus, may improve our results even further. We leave such improvements for future research.

Table 3. *Results: top-15 MWEs*

Hebrew	Gloss	Type
xbr hknst	Member of Parliament	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

the needs of specific applications, and values of k ranging between 2 to 3 have been useful for various applications (Bouma 2009). We conducted experiments with $k = 0.1, 0.2, \ldots, 2.9, 3$ and found $k = 2.7$ to give the best results for our application, maximizing the f-score on the test set. Interestingly, about 15,000 (approximately 10%) of the 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 5). 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-words in the original form (i.e., they are single words split by tokenization).

Refer back to Example 2. 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 hence not considered a candidate. We also consider *adm k*, whose score is low; this prevents the consideration of longer n-gram candidates that include the bigram *adm k* as a substring. Note that the same aligned sentence can be used to induce the *English* MWE *keep away*, which is aligned to a single Hebrew word.

4.6 Results

As an example of the results obtained with this setup, we list in Table 3 the fifteen 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, ninety-nine are

Table 4. *Some results from the top-ranking 100 MWEs*

MWE	Construction
mzg awir (*temper-of air*) "weather"	N+N
kmw kn (*like thus*) "furthermore"	P+ADV
bit spr (*house-of book*) "school"	N+N
šdh t'wph (*field-of flying*) "airport"	N+N
tšwmt lb (*input-of heart*) "attention"	N+N
ai apšr (*not possible*) "impossible"	Particle+ADV
b'l ph (*in-on mouth*) "orally"	P+P+N
ba lidi biTwi (*came to-the-hands-of expression*) "was expressed"	V+P+N
xzr 'l 'cmw (*returned on itself*) "recurred"	V+P+Pron
ixd 'm zat (*together with it*) "in addition"	ADV+P+Pron
h'crt hkllit šl haw"m "the general assembly of the UN"	N+ADJ+P+PN

clearly MWEs.[4] We list some interesting examples, including longer sequences of tokens, in Table 4.

A more careful analysis of the results shows the following pattern. Of the top 1,000 extracted MWEs (of length 2 only), 121 turn out to be false positives (see an analysis of these errors in Section 5.4). Then 266 of the results are proper names: 184 person names, forty-nine geographical terms, and thirty-three miscellaneous names. Recall that the problem of named entity recognition is harder in Hebrew than in European languages; while many of the proper names we extract may have been identified using other means, we view this outcome as an evidence of the robustness of our system. Furthermore, our results include named entities that would have been hard to identify using simple methods such as harvesting Wikipedia. These include *anšil ppr* "Anshil Pepper", an Israeli reporter; and two non-standard spellings of *Ahmet Davutoğlu*.

But our results also include many MWEs that are of very different types. For example, the top-1,000 list includes 262 instances of noun–noun constructions; forty-seven verb–preposition constructions; ninety-seven noun–adjective pairs; fifty-four complex adverbs; nineteen complex conjunctions; etc.

5 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 (two types of intrinsic evaluation, and an extrinsic evaluation) that we detail in this section.

5.1 Intrinsic evaluation

Ideally, one should evaluate the accuracy of a MWE extraction system against a balanced, carefully designed corpus of positive and negative examples, measuring

[4] This was determined by two annotators.

Table 5. *Evaluation results, noun–noun compounds, at a threshold reflecting 2,750 results.*

	TP*	FP†	Precision	Recall	*f*-score
Bigrams (parallel stats)	13	2	0.87	0.11	0.19
Giza++ (parallel stats)	27	3	0.90	0.22	0.36
Giza++ (monolingual stats)	37	8	0.82	0.31	0.45
Bigrams (monolingual stats)	59	15	0.80	0.49	0.61
MWE	67	16	0.81	0.55	**0.66**

*TP: raw number of true positives; †FP: number of false positives.

both precision and recall of the system. Such corpora are of course very difficult to obtain. We were able to obtain a small set of positive and negative MWE instances in a single and specific (albeit frequent) construction. This is the annotated corpus of Hebrew noun–noun constructions (Al-Haj and Wintner 2010), consisting 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, ninety-one 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 $PMI^{2.7}$ to rank the bi-grams in the parallel corpus; using $PMI^{2.7}$ computed from the monolingual corpus to rank the bi-grams in the parallel corpus; and using Giza++ 1:*n* alignments, ranked by their $PMI^{2.7}$ (with bi-gram 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 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 1; the raw data for a threshold reflecting 2,750 results are listed in Table 5.

The plots show the *f*-score of five methods for extracting MWEs, computed on different (but increasingly larger) sets of results. Each column corresponds to a particular (increasingly lower) threshold; of course, the lower the threshold, the more candidates are selected as MWEs, thereby improving the recall but potentially harming the precision. As the graph clearly shows, our results ('MWE') are consistently higher than all other baselines. The difference between the MWE curve and its nearest neighbor, obtained by ranking candidates based on their

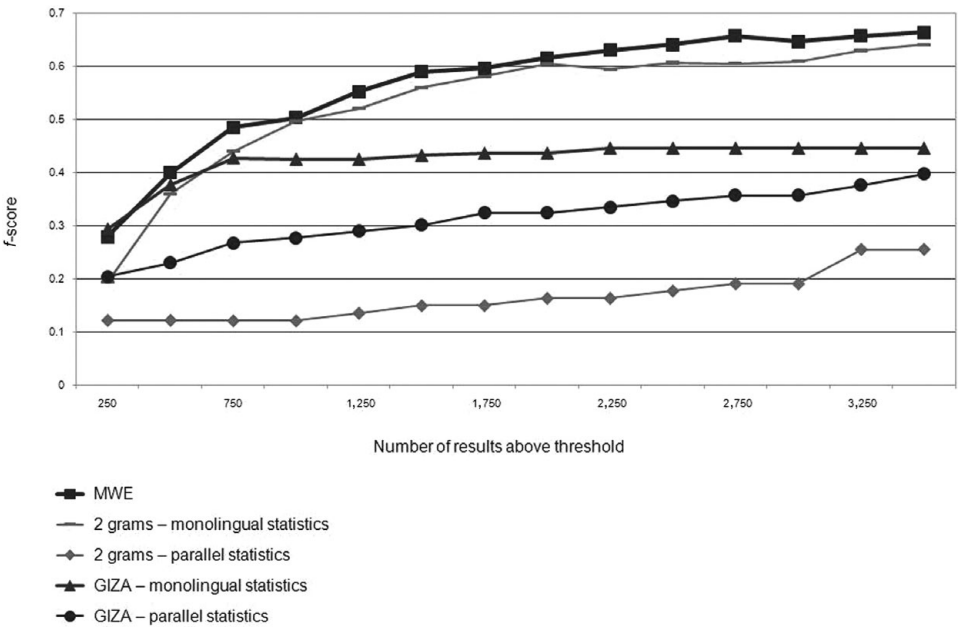

Fig. 1. Evaluation results compared with baselines: noun–noun compounds.

PMI score only, without word alignment, is statistically significant.[5] Specifically, we obtain an f-score of 0.66 at a threshold reflecting 2,750 (and also 3,500) results. The lowest curve reflects statistics drawn from the parallel corpus; these results are poorest due to the small size of the corpus. Interestingly, if a much larger (monolingual) corpus is used, and only the collocation measure is used to determine the MWE status of bi-grams, the results are dramatically better (reflected by the second highest curve). The two middle curves represent the approach that builds on Giza++ alignments, ranked by their PMI score (computed from the parallel and the monolingual corpora, respectively). As can be clearly seen in Figure 1, these curves climb fast but reach a ceiling soon, and setting the threshold lower does not yield much higher f-scores. The best-performing method, no matter where the threshold is set, is our proposed approach.

The advantage of the above evaluation is that it reports both precision and recall of our system. However, these are only measured for a single and very specific construction. To further assess the contribution of our system, we extend the evaluation to other constructions below. However, for arbitrary constructions we only have lists of (positive) MWEs to evaluate against.

For the following experiments, 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,

[5] There can be several ways to determine whether the difference between the two graphs is significant. Assuming that the data are normally distributed, Student's paired t-test shows a confidence level of $p = 0.000011$; but since this is not necessarily the case here, we also compute the Wilcoxon signed-rank test, which yields a confidence level of $p = 0.001$).

Table 6. *Recall evaluation*

	PN		Phrases		NN	
Method	#	%	#	%	#	%
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

Phrases, consists of 571 entries, beginning with the letter *x* in the Hebrew Phrase Dictionary of Rosenthal (2009), and a set of 331 idioms we collected from Internet resources. This set includes arbitrary expressions of various lengths and syntactic constructions, most of which are idiomatic. Of those, 154 occur in the corpus. The third set, **NN**, consists of positive examples in the annotated corpus of noun–noun constructions described above. All instances in this set have the same syntactic structure and similar (high) frequency.

Since we do not have negative examples for two of these sets, we only evaluate recall, using a threshold reflecting 2,750 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 6 the recall of our method (MWE) to Giza++ alignments, as above, and also list the upper bound (UB), obtained by taking all above-threshold bi-grams in the corpus.

The bottom row of Table 6 reflects recall results obtained by focusing on Giza++ $m : n$ alignments, as done in previous works. This approach is likely to miss many proper names, which tend to be 1:1 aligned; hence the poor performance of this method on the **PN** set. As demonstrated above, our methodology is capable of extracting proper names of various types; this is because we validate 1:1 alignments in a dictionary, and many proper names fail this test. The set **Phrases** includes MWEs of various syntactic constructions (and various degrees of semantic opacity), and our approach can clearly identify many of them. Our results are less impressive on the set **NN**, most probably because members of this set are all of the same syntactic construction. The distinction between noun–noun constructions that are MWEs (Al-Haj and Wintner 2010 refer to them as *noun compounds*) and those that are not is not easy, and must rely on several factors, including the semantics of the phrase but also several morphological aspects of its behavior. Our methodology capitalizes on alignment mismatches that are more characteristic of other data sets, and may or may not be able to make the subtle distinctions required for the set **NN**.

5.2 The importance of pre-processing

To emphasize the importance of pre-processing (Section 4.3), we report in this section the results of running exactly the same experiments, *without* first pre-processing the corpora.

The main effect of the lack of pre-processing is during the word-alignment phase, where most of the 1:1 alignments are lost due to data sparsity. Specifically,

Table 7. *Recall evaluation*

Method	PN		Phrases		NN	
	#	%	#	%	#	%
With pre-processing	66	89.2	35	87.5	67	75.3
Without pre-processing	28	37.8	14	35.0	2	2.2

whereas our pre-processed parallel corpus yielded 102,261 1:1 aligned "words" (more precisely, base forms) that are included in our dictionary, without pre-processing, this number sank to 10,886 (a little over 10%). Consequently, the problem of identifying MWEs is reduced to not much more than ranking *n*-grams in a large monolingual corpus (again, with no pre-processing). As a result, most of the extracted MWEs are proper names (which do not tend to be inflected).

We repeated the recall evaluation described above in this setup. We report the results in Table 7, comparing with our previous results. The contribution of pre-processing is evident.

5.3 Extrinsic evaluation

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 frequently wrong). We then associate each extracted MWE in Hebrew with all its possible English translations.

The result is a bilingual dictionary containing 3,750 MWE translation pairs, which we use in the training of a phrase-based Hebrew to English statistical machine translation (SMT) system, exploring its contribution to the quality of the translation, as measured by BLEU (Papineni *et al.* 2002). Specifically, our system is implemented using Moses (Koehn *et al.* 2007), a toolkit for constructing SMT systems. We use our own parallel corpus of approximately 20,000 sentences to train a translation model, and a large monolingual corpus of English newspaper-type texts (obtained from the English Gigaword corpus, Graff and Cieri 2007) for the language model. We randomly selected a set of 1,000 sentence pairs (disjoint from the training set) for tuning and a randomly selected disjoint set of 1,000 sentences for evaluation.

We experiment with three different scenarios of incorporating the MWE dictionary, and compare them with a baseline system, in which the dictionary is not used at all (this is practically the same system that was used by Lembersky, Ordan and Wintner 2011). First, the top-ranking 1,000 Hebrew MWEs, along with their translations, are added to the parallel corpus on which the translation model is based. Second, we use *all* the MWEs extracted by our system (along with their

Table 8. *Contribution to machine translation*

Setup	BLEU
Baseline	12.89
Top-ranking MWEs	13.02
All MWEs	**13.47**
All MWEs, upweighted	13.43

translations, of course), and finally we use all MWEs again, but we duplicate them three times in order to upweight the MWEs compared with the default training material. This is similar to the evaluation technique of Carpuat and Diab (2010).

The results are depicted in Table 8. In all cases, incorporating MWEs in the system results in an improved BLEU score. The best system, in which all MWEs are added to the training material, significantly improves the baseline ($p = 0.022$). Some examples of improved translations with the best performing system include *the health system* (compare with *the health*, generated by the baseline system); *the nation state was founded* (vs. *the nation state was*); *the october events* (vs. *the events of the past october*); *for the sake of plans and a short-term proposition* (vs. *for the sake of plans a long-term short*); and *the government of prime minister benjamin netanyahu* (vs. *minister benjamin netanyahu government*).

5.4 Error analysis

Our MWE extraction algorithm works as follows: Translated texts are first sentence-aligned. Then, Giza++ is used to extract 1-to-1 word alignments, that are then verified by the dictionary and replaced by '*' if the correct word translation is available. This process filters out candidates that have compositional meaning and, therefore, are not considered MWEs (in our algorithm, a non-compositional meaning of a bi-gram is expressed by its non-literal translation to the parallel language). Sequences of words separated by '*'s are considered MWE candidates. At each step of the application errors may occur that lead to false identification of non-MWEs. We manually annotated the top 1,000 bi-gram MWEs extracted by the algorithm and identified 121 false positives. Analysis of these false positives reveals the error sources detailed below. In Table 9 we summarize the statistics of the error sources.

Translation quality of the parallel corpus. Whereas the sentences are indeed translations, the translations are, to a large extent, non-lexical in the sense that context is used in order to extract the meaning and deliver it in different wording. As a result, it is sometimes hard or even impossible to align words based on the sentence alone.

As an example, a newspaper text includes, on the English side, a sentence beginning with a reported utterance, followed by *according to senior officials*. Its Hebrew translation uses *kk msrw pqidim bkirim* (*thus reported officials senior*) "said senior officials." As a result, Hebrew *kk msrw* "thus reported" is aligned with *according*, and is considered a MWE candidate.

Table 9. *Sources of errors*

Error source	False positives	
	#	%
Translation quality of the parallel corpus	46	38.02
Sentence alignment errors	19	15.70
Word alignment errors	21	17.36
Noise introduced by pre-processing	29	23.97
Incomplete dictionary	4	3.31
Parameters of the algorithm	2	1.65

Sentence alignment errors. Several errors can be attributed to the automatic sentence alignment.

(1) We use a purely statistical sentence aligner to align sentences based on their length and token co-occurrence information. As a result, some sentences of similar length may incorrectly be marked as mutual translations. Of course, most of the word sequences in such sentences cannot be aligned and hence become MWE candidates.

(2) The output of the sentence aligner contains only 1-to-1 sentence translations. As our parallel corpora include non-lexical translations that sometimes can only be expressed in terms of 1-to-2, or 2-to-1 translated sentences, the sentence aligner may output a 1-to-1 alignment, where one of the sentences is only a partial translation of another. The non-translated part of the sentence may contain false MWE candidates.

Word alignment errors. Sometimes a word sequence has a translation, but it is not aligned properly. Possible reasons for such errors are as follows:

(1) Insufficient statistics of word co-occurrence due to the small size of the parallel corpus.

(2) Errors caused by bi-directional translation merge (we employ union to merge the translations in both directions (Och and Ney 2003); intersection resulted in worse results). Often the alignment is correct only in one direction, but we lose this information after merging the alignment; this often happens in very long sentences. Another example of the problematic alignment caused by bi-directional merge is cases in which the word aligner proposes n:1 alignment; usually these n words contain the correct sequence or a part of the sequence and the correct analysis of the bi-directional alignments may help filter out the incorrect parts (i.e., the analysis of the intersection of n and m sequences, where m:1 is Hebrew-to-English and n:1 is English-to-Hebrew alignments detected by the word alignment tool).

As an example, our Hebrew corpus includes the sentence *lm'šh, drwš lw rq kšrwn axd, lškn' anšim lhcbi' b'dw* (*in-fact, required to-him only talent one, to-convince people to-vote for-him*) "in fact, he only needs one talent: to convince the electorate to vote for him". This is aligned against the English

He needs only one talent: to convince the electorate to vote for him. Giza++, however, aligns the Hebrew *lhcbi' b'dw* "to-vote for-him" with the English *vote*, whereas the final English *him* is aligned with the Hebrew third token (*lw* "to-him"), and the English penultimate *for* is aligned to *null*.

Noise introduced by pre-processing

(1) Errors caused by morphological analysis and disambiguation tools may lead to wrong tokenization, or to the extraction of an incorrect base form from the surface form of the word. As a result, the extracted citation form cannot be aligned to its translation, and correctly aligned word-pairs cannot be found in the dictionary. For example, the bi-gram *bniit gdr* (a noun–noun compound) is translated as *building fence*. Stemming on the English side produces the erroneous base form *build* (a verb) for the word *building*. Word alignment correctly aligns the words *bniih* (a noun) and *build* (a verb), but such a pair does not exist in the dictionary, which contains the following pairs: *bnh-build* (verb), and *bniih-building* (noun).

(2) An additional source of errors stems from language-specific differences in word order between the languages: e.g., *txnt rkbt* is consistently translated as *railway station*; the correct alignment would be *txnt—station*, *rkbt—railway*, but due to the different word order in the two languages, and to the fact that both phrases are frequent collocations, Giza++ proposes the alignment *txnt—railway*, *rkbt—station* (these pairs are not in the dictionary and, therefore, the bigram *txnt rkbt* is falsely identified as an MWE). Such problems can be handled with more sophisticated pre-processing that reduces language-specific differences, where not only morphology and function words are taken into account but also language-specific word order.

Incomplete dictionary. If sentence and word alignments are correct, and the correct word-to-word translation exists, but the translated pair is not in the dictionary, the word sequence may erroneously be considered an MWE candidate.

Unfortunately, we have no remedy for most of these errors, other than using larger corpora and better language resources. This is the kind of noise that is likely to affect any natural language processing application.

Parameters of the algorithm

(1) Setting the threshold too high causes bi-grams that are subsequences of longer MWEs to be false positives. For example, the non-MWE, compositional bi-gram *lšlm ms* "pay tax", which is a subsequence of the MWE *lšlm ms šptiim* (*pay tax-of lip*) "pay lip service", was mistakenly extracted as an MWE, since the score of the bi-gram *ms šptiim* "lip tax" is lower than the threshold.

(2) During error analysis we revealed the following algorithm drawback (which is probably common to other alignment-based methods): False MWE candidates that occur several times in the parallel corpus are selected to be MWE candidates only in a minority of these occurrences. In other words,

we define as MWE candidate any *n*-gram that was misaligned; we do not check whether this *n*-gram was misaligned consistently in all (or most) of its occurrences in the corpus. For example, there are twelve occurrences of the bi-gram *nšia hmdinh* (president of the state) in the parallel corpus, but only twice does it appear as a candidate bi-gram due to two sentences in which the translation of this bi-gram is missing (due to non-literal or incorrect sentence translation). From this we conclude that the algorithm can also be improved if the candidates would be selected from bi-grams that have no translation in the parallel language in a majority of their occurrences. We leave this improvement for future work.

6 Conclusions and future work

We described a methodology for extracting multi-word 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 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 use 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 demonstrate 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. Furthermore, out methodology is basically language-independent. Indeed, we applied the same approach to a large English–French parallel corpus, consisting of nearly five million words. While we do not have a way to properly evaluate the results, the top candidates in both languages were all clearly MWEs. We therefore believe that this methodology is easily applicable to other language pairs for which a small parallel corpus (and a larger monolingual one) exist. Finally, as our main motivation is high-precision, high-recall extraction of Hebrew MWEs, we would like to explore the utility of combining different approaches to the same task (Al-Haj and Wintner 2010) under a unified framework. A first attempt in this direction is reported in Tsvetkov and Wintner (2011).

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 invaluable help with the machine translation system.

References

Al-Haj, H. February 2010. *Hebrew Multiword Expressions: Linguistic Properties, Lexical Representation, Morphological Processing, and Automatic Acquisition.* Master's thesis, University of Haifa, Haifa, Israel.

Al-Haj, H., and Wintner, S. 2010. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, August 2010, pp. 10–18. http://www.aclweb.org/anthology/C10-1002

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword expressions*, Sapporo, Japan, pp. 89–96. Stroudsburg, PA, USA: Association for Computational Linguistics.

Baldwin, T., and Tanaka, T. July 2004. Translation by machine of complex nominals: getting it right. In T. Tanaka, A. Villavicencio, F. Bond, and A. Korhonen (eds.), *Second ACL Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain, pp. 24–31. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bannard, C, Baldwin, T, and Lascarides, A. 2003. A statistical approach to the semantics of verb-particles. In D. M. F. Bond, A. Korhonen, and A. Villavicencio (eds.), *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 65–72. http://www.aclweb.org/anthology/W03-1809.pdf

Bar-Haim, R., Sima'an, K., and Winter, Y. June 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*, Ann Arbor, MI, USA, pp. 39–46. Stroudsburg, PA, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W05/W05-0706

Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python.* Sebastopol, CA: O'Reilly Media.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Von der Form zur Bedeutung: Texte Automatisch Verarbeiten/From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, Tübingen: Gunter Narr Verlag, pp. 31–40.

Brants, T., and Franz, A. 2006. Web 1T 5-gram version 1.1. LDC Catalog No. LDC2006T13. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

Carpuat, M., and Diab, M. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, CA, USA, June 2010, pp. 242–5. Stroudsburg, PA, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/N10-1029

Caseli, H., Villavicencio, A., Machado, A., and Finatto, M. J. 2009. Statistically driven alignment-based multiword expression identification for technical domains. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, Singapore, August 2009, pp. 1–8. Stroudsburg, PA, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W09/W09-2901

Chang, B., Danielsson, P., and Teubert, W. 2002. Extraction of translation unit from Chinese-English parallel corpora. In *Proceedings of the first SIGHAN Workshop on Chinese Language Processing*, Morristown, NJ, USA, pp. 1–5. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dx.doi.org/10.3115/1118824.1118825

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* **16**(1):22–9. ISSN 0891-2017.

Cook, P., Fazly, A., and Stevenson, S. 2007. Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions (MWE 2007)*, Prague, Czech Republic, June 2007, pp. 41–8. Stroudsburg, PA, USA: ACL.

Daille, B. 1994. *Approche Mixte Pour L'extraction Automatique de Terminologie: Statistiques Lexicales et Filtres Linguistiques*. PhD thesis, Université Paris, Paris, France.

Dejean, H., Gaussier, E., Goutte, C., and Yamada, K. 2003. Reducing parameter space for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts*, Morristown, NJ, USA, pp. 23–6. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dx.doi.org/10.3115/1118905.1118910.

Doucet, A., and Ahonen-Myka, H. 2004. Non-contiguous word sequences for information retrieval. In T. Tanaka, A. Villavicencio, F. Bond, and A. Korhonen (eds.), *Second ACL Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain, July 2004, pp. 88–95. Stroudsburg, PA, USA: Association for Computational Linguistics.

Erman, B., and Warren, B. 2000. The idiom principle and the open choice principle. *Text* **20** (1):29–62.

Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: Language, Speech and Communication, MIT Press.

Graff, D., and Cieri, C. 2007. *English Gigaword*, 3rd. ed. LDC Catalog No. LDC2007T07. Philadelphia, PA, USA: Linguistic Data Consortium.

Itai, A., and Wintner, S. March 2008. Language resources for Hebrew. *Language Resources and Evaluation* **42**(1):75–98.

Jackendoff, R. 1997. *The Architecture of the Language Faculty*. Cambridge, MA, USA: MIT Press.

Katz, G., and Giesbrecht, E. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, July 2006, pp. 12–19. Stroudsburg, PA, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W06/W06-1203

Kirschenbaum, A., and Wintner, S. 2010. A general method for creating a bilingual transliteration dictionary. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010, pp. 273–6. Paris, France: European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

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

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. June 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, June 2007, pp. 177–80. Stroudsburg, PA, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/P07-2045

Lambert, P., and Banchs, R. 2005. Data inferred multi-word expressions for statistical machine translation. In *Proceedings of the MT Summit X*, Phuket, Thailand, pp. 396–403.

Lapata, M., and Keller, F. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing* **2**:1–31.

Lembersky, G., Ordan, N., and Wintner, S. 2011. Language models for machine translation: original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, July 2011, pp. 363–74. Stroudsburg, PA, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/D11-1034

Melamed, I. D. 1997. Measuring semantic entropy. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics*, pp. 41–6.

Nakov, P., and Hearst, M. 2005. Search engine statistics beyond the n-gram: application to noun compound bracketing. In *Proceedings of CoNLL '05*, pp. 17–24. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.9694

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, pp. 311–8. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dx.doi.org/10.3115/1073083.1073135.

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

Piao, S. S., Rayson, P., Archer, D., and McEnery, T. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language* **19**(4):378–97. ISSN 0885-2308. http://dx.doi.org/10.1016/j.csl.2004.11.002.

Ren, Z., Lü, Y., Cao, J., Liu, Q., and Huang, Y. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, Singapore, August 2009, pp. 47–54. Stroudsburg, PA, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W09/W09-2907

Rosenthal, R. 2009. *Milon HaTserufim* (*Dictionary of Hebrew Idioms and Phrases*) (in Hebrew). Jerusalem: Keter.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. 2002. Multiword expressions: a pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics* (*CICLING 2002*), Mexico City, Mexico, pp. 1–15.

Smadja, F. A. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* **19** (1):143–77.

Tsvetkov, Y., and Wintner, S. 2010a. Automatic acquisition of parallel corpora from websites with dynamic content. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, May 2010, pp. 3389–92. Paris, France: European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Tsvetkov, Y., and Wintner, S. 2010b. Extraction of multi-word expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics* (*COLING 2010*), Beijing, China, August 28, 2010.

Tsvetkov, Y., and Wintner, S. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, July 2011, pp. 836–45. Stroudsburg, PA, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/D11-1077

Uchiyama, K., Baldwin, T., and Ishizaki, S. October 2005. Disambiguating Japanese compound verbs. *Computer Speech & Language* **19**(4):497–512.

Van de Cruys, T., and Villada Moirón, B. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on a Broader Perspective on Multiword*

Expressions, Prague, Czech Republic, June 2007, pp. 25–32. Stroudsburg, PA, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W07/W07-1104

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP'2005*, Borovets, Bulgaria, September 21–23, 2005, pp. 590–6.

Venkatapathy, S., and Joshi, A. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, July 2006.

Venkatsubramanyan, S., and Perez-Carballo, J. 2004. Multiword expression filtering for building knowledge. In T. Tanaka, A. Villavicencio, F. Bond, and A. Korhonen (eds.), *Second ACL Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain, July 2004, pp. 40–7. Stroudsburg, PA, USA: Association for Computational Linguistics.

Villada Moirón, B., and Tiedemann, J. 2006. Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the EACL 2006 Workshop on Multi-Word-Expressions in a Multilingual Context*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., and Ramisch, C. 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)*, pp. 1034–43. http://www.aclweb.org/anthology/D/D07/D07-1110

Zarrieß, S., and Kuhn, J. 2009. Exploiting translational correspondences for pattern-independent MWE identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, Singapore, August 2009, pp. 23–30. Stroudsburg, PA, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W09/W09-2904