Translation of Multiword Expressions Using Parallel Suffix Arrays

Paul McNamee and James Mayfield
Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, Maryland 20723-6099 USA
{paul.mcnamee,james.mayfield}@jhuapl.edu

Abstract

Accurately translating multiword expressions is important to obtain good performance in machine translation, cross-language information retrieval, and other multilingual tasks in human language technology. Existing approaches to inducing translation equivalents of multiword units have focused on agglomerating individual words or on aligning words in a statistical machine translation system. We present a different approach based upon information theoretic heuristics and the exact counting of frequencies of occurrence of multiword strings in aligned parallel corpora. We are applying a technique introduced by Yamamoto and Church that uses suffix arrays and longest common prefix arrays. Evaluation of the method in multiple language pairs was performed using bilingual lexicons of domain-specific terminology as a gold standard. We found that performance of 50-70%, as measured by mean reciprocal rank, can be obtained for terms that occur more than 10 or so times.

1 Introduction

When processing human language it is difficult to operate only at the level of individual words. While for some tasks, perhaps monolingual information retrieval in particular, this might seem reasonable, for others such as machine translation, cross-language question answering, and translingual information retrieval, restriction to processing single words is a significant impediment. There has been much recent interest in computational approaches to dealing with multiword expressions (MWEs) as workshops at ACL-2006, SIGIR-2005, ACL-2004 and other conferences attest.

Even identifying what constitutes a good MWE is difficult, and there are a variety of interesting classes such as idiomatic expressions, complex noun phrases, phrasal verbs, and collocations. Many researchers have focused attention on short-length n-grams (notably $n=2$) or phrases that compose a syntactic unit. Longer length n-grams have been eschewed due to computational complexity and data sparseness. Here we examine arbitrary sequences of words and attempt to translate them.

To effect translation we will rely on information-theoretic measures such as Dice scores or Mutual Information, and large aligned bilingual texts. To score translation candidates we will compute global and local frequencies of occurrence of substrings using the method described by Yamamoto and Church (2001) which is based on suffix arrays and longest common prefix arrays.

In the rest of this paper we review earlier work in phrase translation (Section 2) and fully describe our approach (Section 3). In Section 4 details of our evaluation are presented and we conclude the paper with a discussion of our results (Section 5).

2 Related Work

The problem of phrasal translation has been attempted in three different ways: (1) fusion of translations of component words; (2) translation using word alignments and bilingual suffix trees; and (3) purported phrases induced by word alignments in a SMT system. The present work can be seen as a

Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 100-109,
Cambridge, August 2006.

simplification of the second paradigm that does not require word alignments and therefore can use contingency table methods.

2.1 Contingency Table Methods

We first review three efforts to translate multiword units by scoring POS-tagged phrases or by fusing individual target language words that appear correlated to a source language phrase.

Kupiec (1993) examined translation of noun phrases between English and French and reported 90% accuracy in an informal evaluation of the one hundred translations that had the highest confidence scores. His method requires POS-tagging each sentence in an aligned parallel text (*i.e.,* both sides are tagged). Then noun phrases are scored using an iterative estimation method. Kupiec notes several sources of error caused by problems using POS information instead of constituent parses, such as an inability to infer prepositional phrase attachment. The method relies on having POS tagging or parsing in both languages.

Dagan and Church developed, *Termight*, a tool that was meant assist professional translators and terminologists develop bilingual term lists and technical terminology in particular (1997). Like Kupiec's work, they also presume the availability of POS-tagging and work with noun phrases extracted from sentence-aligned corpora. A distinctive feature of their approach is using word-level alignments to score translations; this enables identification of correct translations even with the correct source term / target term correspondence is observed once or twice in the bilingual data. (This scenario, when term frequency is small, makes translation using contingency table methods such as Dice coefficients problematic.) They report finding a correct translation at rank 1 40% of the time and at rank 2 an additional 7% of the time for a list of 192 English/German technical terms.

The *Champollion* system was developed by Smadja et al. (1996) to specifically address translation of collocations, including non-compositional expressions. They used aligned sentences from the Canadian Hansards corpus (as did Kupiec). They used a tool they had previously developed (XTRACT) to identify collocations and they translated approximately 900 medium frequency English phrases to French. Manual evaluation by bilingual speakers revealed accuracies between 65 and 78%. Champollion works by iteratively fusing together target language words that are strongly correlated to the source language collocation for which translation is attempted. Dice scores are used to filter out unlikely word combinations.

2.2 Bilingual Suffix Trees

Munteanu and Marcu (2002) ambitiously produce alignments from comparable corpora, corpora where exact translations may not be available, but in which the same topics or entities are being discussed such as contemporaneous newswire. They use suffix trees in both languages and a bilingual lexicon to provide points of correspondence between the two languages. Not only to they successfully create alignments in the comparable data, thus creating a parallel corpus, they create phrasal alignments of a restricted sort. Namely, their parallel phrases have the same number of tokens (*i.e.,* words) in each language and the word order of the source and target languages must be the same. Some examples of English/French alignments that they identified are:

- mon avis, ce est très important / my opinion, this is very important
- par mes collègues et moi / by my colleagues and myself
- toutes les personnes / all the people

They estimated that 95% of their aligned sequences were correct. They reported that it took only about 1.5 minutes to build the suffix trees, but between 38 and 60 hours to create matches between the two suffix trees and extract contiguous phrases. While effective for creating alignments in comparable corpora, this approach suffers from the restriction of working with languages having a common word order and only being able to produce isometric phrases.

2.3 Phrase-Based SMT

Recently researchers in statistical machine translation have developed methods to move beyond single word alignments and create richer translation models that contain phrasal alignments (Och and Ney, 2004). The general method is to induce single word alignments using maximum likelihood estimates obtained from parallel data such as by IBM

Model 1 (Brown *et al.,* 1993) and to use these alignments to suggest adjacent words that may compose a meaningful phrase. By examining bidirectional alignments for the same parallel data a 'symmetrized alignment matrix' can be obtained, and from this information potential translations of word sequences can be obtained. As long as contiguous sequences are examined it does not matter if the two languages have different word order. The approach can be further generalized by working with word classes so that hypotheses for unseen phrases can be generated.

A few researchers have started to exploit suffix arrays in phrase-based SMT systems. Callison-Burch *et al.* (2005) have shown how parallel suffix arrays can be used to efficiently compute phrase translations and significantly reduce the large memory footprints that phrased-based SMT systems suffer from when attempting to use longer (*i.e., n>3*) phrases. Zhang and Vogel (2005) describe a dynamic programming algorithm that more efficiently retrieves alignments for a set of phrases (such as all substrings from a sentence that is to be translated), which outperforms direct comparison binary search by a couple of orders of magnitude. Their improvement allows them to compute phrase alignments online.

The present work is distinct from these techniques in that it does not depend on iteratively trained word alignments to postulate phrasal translations and thus has an advantage in reduced computational expense. The algorithm can be run online or it can be used to precompute phrase translation tables for all n-grams.

3 Translation of Arbitrary Phrases

Suffix arrays were introduced by Manber and Myers who gave a Θ(N log N) construction algorithm (1991). While several linear time suffix array construction algorithms have now been introduced (Kärkkäinen and Sanders, 2003; Ko and Aluru. 2003) it is not clear that their asymptotic gains make them a better choice than well-tuned supralinear methods on corpora of interest (Puglisi *et al.,* 2005).

3.1 Yamamoto/Church Algorithm

Irrespective of how the suffix array is created, Yamamoto and Church (2001) demonstrated how given a suffix array, frequencies of occurrence for all substrings can be ascertained in linear time. More precisely, by doing O(N) preprocessing, frequency information about any substring can be obtained in O(log N) time. Enumerating over all substrings naturally requires quadratic time. However, their technique works by partitioning substrings into at most 2N classes which have unique *collection frequency*, the number of times the string occurs in the text, and *document frequency*, the number of separate documents the string occurs in. (The text may contain special end-of-document markers.)

N simple classes exist, each corresponding to a distinct suffix of the text. The remaining classes are created by examining the longest common prefix array and identifying intervals where LCP values inside the interval are greater than the surrounding values. These non-trivial classes of suffixes correspond to sets of substrings with prefixes in common. From these common prefixes we identify translation candidates.

In Yamamoto and Church's work, they demonstrate how interesting multiword phrases can be discovered; by interesting they mean phrases that may be beneficial for information retrieval or computational lexicography as determined by Mutual Information (MI) or Residual Inverse Document Frequency (RIDF) (Church, 1995) scores. Here we are concerned with scoring candidate phrase translations.

3.2 Contingency Table Methods

Various information-theoretic scoring methods can be used to measure the correlation of two terms. Dice coefficients and Mutual Information are commonly used in the field of corpus linguistics to discover meaningful term associations. Each can be computed using a contingency table that stores the frequencies of two terms occurring separately and together. Such tables naturally support maximum likelihood estimates of: $P(x,y)$, $P(\bar{x},y)$, $P(x,\bar{y})$, and $P(\bar{x},\bar{y})$ for two terms *x* and *y*. When studying monolingual corpora and searching for term associations, $P(x,y)$ is based on the frequency that x and y co-occur together in the same number of contexts (*i.e.,* documents). To search for potential translations of a term x using aligned bilingual corpora, we will interpret $f(x,y)$ to be the number of times that y is found in documents that

correspond to documents in which x appeared. Note, unlike the monolingual case, this interpretation need not be symmetric as $f(y,x)$ may not equal $f(x,y)$. For example, suppose in parallel English/French data, automobile (EN) is translated alternatively as automobile or voiture, and voiture (FR) can be rendered as vehicle or automobile.

The method proposed here is analogous to isolated word translation using contingency tables except we are working with arbitrary length expressions. We measure $f(y,x)$ using exact counts; however, to avoid quadratic behavior we do not enumerate every possible substring.

3.3 Scoring Candidate Translations

Our algorithm starts by taking aligned corpora and constructing suffix arrays for both the source and target language texts. We retain these suffix arrays in memory and also store with each a much smaller array that identifies document boundaries. Now, for a word or phrase that we wish to translate, we perform binary search to identify documents (or sentences) containing the input. The corresponding target language documents for each source document containing the input are concatenated into a new string, T, which represents a subset of the target language collection. (For expediency, if the input phrase is very common we cap the number of documents at 10,000). From T a third suffix array, SA_3, is now constructed.

We apply Yamamoto/Church to this new suffix array to partition substrings into classes. A Dice score for each class (*i.e.,* set of substrings with common prefix) is computed. The joint frequency is taken from the subcollection using SA_3, and the individual frequencies of occurrence are taken from the larger source and target language suffix arrays. Each scored class is added to a priority queue and the *k*-best candidates are returned.

There is one complication that we have glossed over in the previous description. If a class contains more than one string each string should be scored. But there could be a large number of strings in a class. For example, if a source language term only appears in a single sentence of length 20 words, there could be $O(20^2)$ substrings to consider. This could become expensive, so the following heuristic is used: if there are fewer than 8 sentences in T, consider all substrings. When more target language sentences are available, only a single string from each class is selected - the longest prefix.

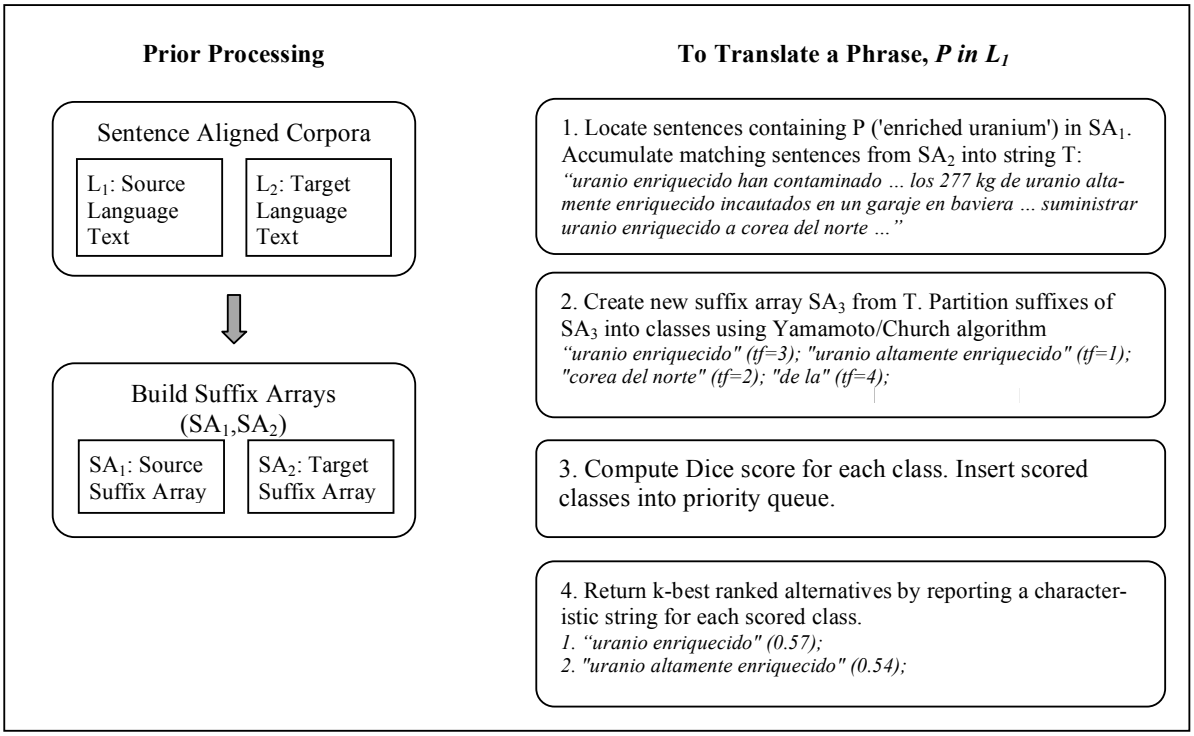

Figure 1. Translation Using Parallel Suffix Arrays

For the more frequently occurring terms the above heuristic works well, but for terms that only occur in 1 or 2 sentences, it is difficult to extract the correct string without some other knowledge (*i.e.,* part-of-speech information; a bidict or word alignments; similar spellings in related languages). To further illustrate this point, suppose the only sentence under consideration was: "the quick brown fox jumped over the lazy dog." Substrings like:

- 'brown fox jumped'
- 'brown fox jumped over
- 'brown fox jumped over the lazy dog'

will each be considered equally valid translations of the source language term.

Sample translations using the method are given in Table 1.

Table 1. Sample translations

English Term	Freq	Spanish Term	Dice Score
enriched uranium	7	uranio enriquecido	0.571
		altamente enriquecido	0.546
		uranio altamente enriquecido	0.546
first of all I should like	313	primer lugar quisiera	0.140
		en primer lugar quisiera	0.140
		lugar quisiera	0.121
foot and mouth disease	371	aftosa	0.723
		fiebre aftosa	0.720
		fiebre	0.702

While we have described the algorithm from the perspective of searching for translations of a specified input, a translation table can be created for all phrases. We have done this by again using the Yamamoto/Church algorithm. We simply partition the source language text into classes and produce candidates for each. In this way we obtained translations for every substring occurring 10 or more times in our data in about a CPU-week using a non-optimized implementation.

4 Experiments

We require aligned parallel text to fuel our methods and we relied on the Europarl corpus (Koehn, 2003) which is comprised of EU parliamentary oration that has been manually translated into other EU languages. There is approximately 160 MB of text (about 24 million words) per language.

To verify the efficacy of our proposed method on a variety of language pairs we decided to use bilingual wordlists to measure successful translation. We sought lists of technical terminology that we hoped would contain many MWEs. Table 2 lists the resources we relied upon. Four separate dictionaries were used covering English (EN) and five other Western European languages: German (DE), Spanish (ES), French (FR), Italian (IT), and Portuguese (PT). Some of the dictionaries we found were truly multilingual with alignments possible between any language pair, but others were only available where English was one of the languages involved. All of our experiments in this study use English as either the source or target language. In Table 2 the number of dictionary entries is given both for all entries and for those containing a multiword expression on either the source or target side.

Typically fewer than 10% of these terms are actually present in the Europarl corpus and we evaluate performance only when the source language entry appears at least once in the data (regardless of whether a correct target language entry occurs as well). In the table the number of terms that appearing in the English side of our aligned corpus is given in parentheses.

Table 2. Wordlists used

	Source	Langs.	Domain	All Pairs	MWE Pairs
FOOD	todine.net[1]	EN DE ES FR IT	Culinary terms	527 (84)	421 (32)
IFCC	ifcc.org[2]	EN ES	Science	2811 (529)	1600 (103)
IMF	International Monetary Fund[3]	EN DE ES FR PT	Finance	5284 (1822)	4977 (1575)
UN	United Nations[4]	EN DE ES FR IT	Product codes	14572 (1173)	13213 (583)

A candidate translation was deemed correct when it exactly matched a correct translation from the bilingual dictionaries that serve as gold standards. No attempt is made to normalize the dic-

[1] http://www.todine.net/dictionary.html
[2] http://www.ifcc.org/divisions/CPD/dict/spandict.htm
[3] http://www.imf.org/external/np/term/index.asp
[4] http://www.unspsc.org/ (version 5.0301)

tionaries for mistakes in spelling, dialectal variation (*i.e.,* color_television vs. colour_television), plural forms or gender variations, and the presence or absence of diacritical marks. In one of the French dictionaries we removed leading articles (*e.g.,* les) that did not appear to correspond to the English entries. We did remove punctuation (not hyphens) and perform case normalization of both the parallel text and our bilingual terms. Our strict criterion for translation correctness removes subjectivity, but it slightly depresses performance.

We computed several metrics to ascertain performance: namely, the percentage of time that a correct answer is found at rank 1; no higher than rank 3; and, the mean reciprocal rank (MRR) of the first accurate translation. The inverse of MRR is the expected rank of a valid translation. While precision of the first response is an intuitive measurement, end applications such as cross-language information retrieval or phrase tables in a SMT system's translation model can make use of multiple translation alternatives, thus we wanted to measure MRR or precision at a slightly deeper depth. We computed these metrics both in aggregate and for subsets of source language terms partitioned according to the term frequency of the source term. Translation accuracy as a function of source term corpus frequency is plotted in Figure 2. As described previously we used Dice scores and we considered up to 25 candidate translations.

Performance is plotted on the vertical axis and measured for different frequency bins. The Zipfian distribution of terms suggests assessing performance by frequency ranges growing by a constant factor (Zipf, 1949); accordingly we used a logarithmic scale with base=3. The number of terms per bin is given in parentheses. Some of the previous work cited in Section 2 described performance for selected subsets of terms, typically high frequency terms that are easier to translate. We believe presenting translation accuracy as a function of source term frequency is more informative.

Examining the figure it is apparent that the three curves are in strong agreement. MRR scores are naturally higher than P@1, but not much lower than P@3. Performance rises quickly with term frequency and for terms that occur about 10 or more times in our parallel data, both MRR and P@3 are between 50% and 70%. P@1 trails by approximately 5%.

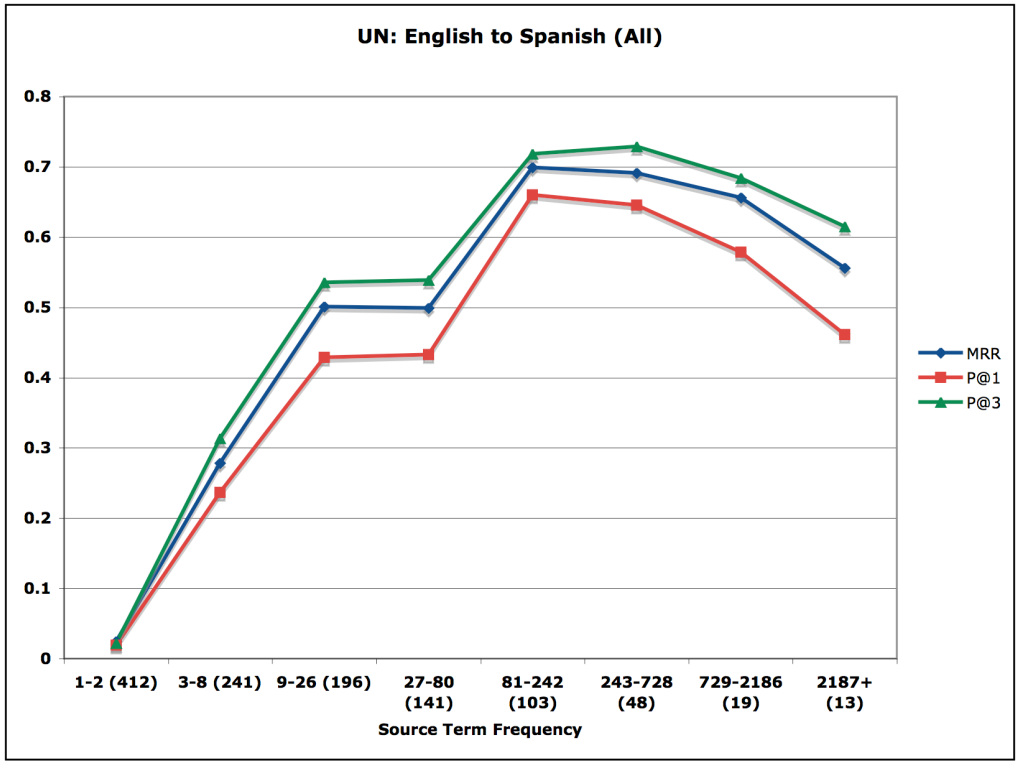

Figure 2. Translation effectiveness computed over all terms.

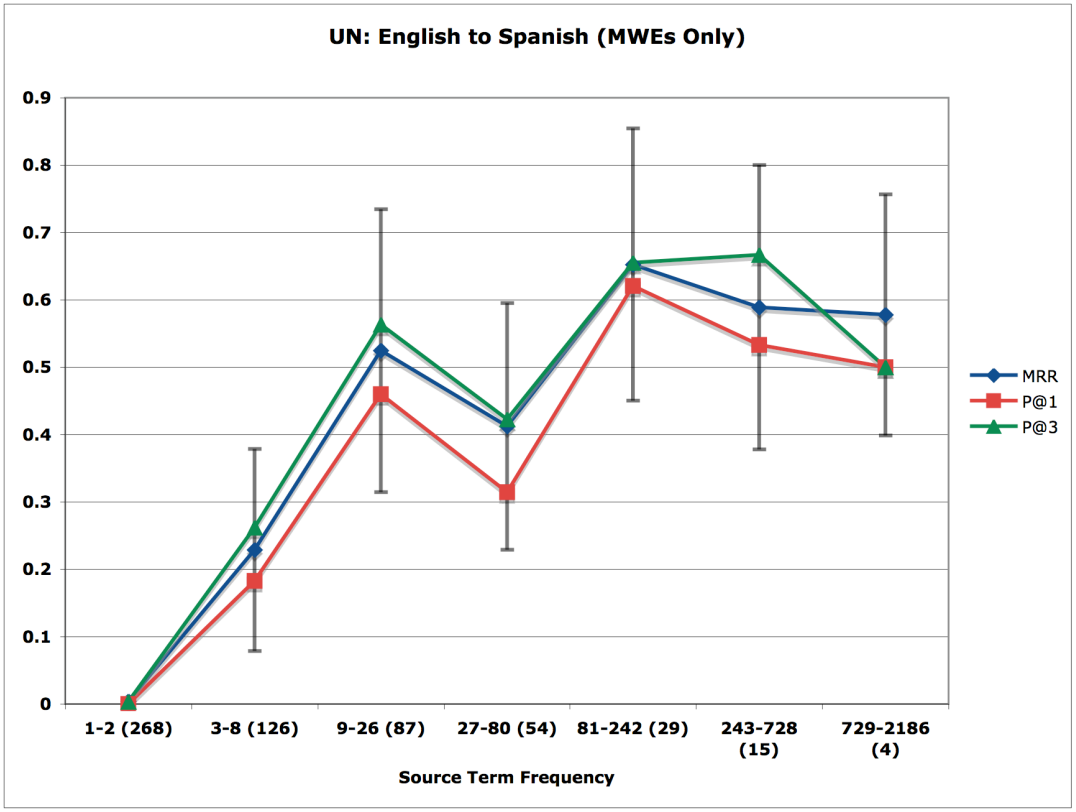

Figure 3. Translation effectiveness computed only for MWEs.

Although we targeted bilingual dictionaries in diverse domains to increase the number of MWEs available for translation, the data in Figure 2 is based on all source terms in the bilingual dictionary that were present in the data, and many are single words. To better quantify MWE translation accuracy we also examined performance when a source language term or a target language translation of that term was made up of multiple words. These results are presented in Figure 3.

Here, as in Figure 2, performance rapidly improves for more frequent terms. MWE performance is slightly lower on average than with individual words and corresponding points in Figure 3 are about 5% lower than in Figure 2, so. In fact, this situation is more evident from examining the data in Table 3 (below) – a drop of 10-20% often occurs, but not with the IMF dataset.

There is a knee in the curves in Figure 3 at the 27-80 bin, which we initially ascribed to variance; however as the bars on the graph show, the variance in MRR is fairly consistent for all but the lowest frequency terms. We do not have an explanation for the effect.

In Table 3 performance is reported using 13 bilingual dictionaries from four sources. Translation was performed both directions, thus there are 26 cases. Each row in the table has four regions corresponding to: averages when the source term occurred 5 or more times (left half) or for all frequencies, and measured only for both all terms (MWEs and words) and MWEs only. N is the number of source terms that met the criteria. Clearly when lower frequency terms are left out performance is greater because they are harder to translate and also their greater number contributes more to averages than more common terms. Averages were computed across language pairs in each dictionary to suggest the relative difficulty of each dictionary's terms; the FOOD and IFCC data appear to be the easiest.

Table 3. Performance using four bidicts.

		Terms (CF >= 5)			MWEs Only (CF >= 5)			Terms (All Freq.)			MWEs Only (All Freq.)		
Bidict	Pair	N	MRR	P@1	N	MRR	P@1	N	MRR	P@1	N	MRR	P@1
FOOD	ENDE	50	0.5473	0.4800	15	0.3667	0.3333	84	0.3868	0.3452	37	0.2143	0.2000
	DEEN	41	0.5700	0.4634	10	0.4000	0.3000	73	0.3851	0.3151	23	0.2609	0.2174
	ENES	50	0.5930	0.5400	15	0.3556	0.3333	84	0.4032	0.3571	32	0.2146	0.1875
	ESEN	57	0.5416	0.4737	14	0.3929	0.3571	82	0.4294	0.3780	27	0.2407	0.2222
	ENFR	50	0.7195	0.6800	12	0.4444	0.3333	84	0.4439	0.4048	32	0.1667	0.1250
	FREN	60	0.4969	0.4167	16	0.3458	0.2500	85	0.3831	0.3176	26	0.2128	0.1538
	ENIT	50	0.6039	0.5000	12	0.5444	0.3333	84	0.4079	0.3333	32	0.2042	0.1250
	ITEN	58	0.5421	0.4655	18	0.3333	0.2778	106	0.3296	0.2830	48	0.1474	0.1250
Average			**0.5768**	**0.5024**		**0.3979**	**0.3148**		**0.3961**	**0.3417**		**0.2077**	**0.1695**
IFCC	ENES	371	0.5448	0.5040	51	0.2300	0.1961	529	0.4131	0.3819	103	0.1340	0.1165
	ESEN	390	0.5323	0.4795	39	0.3549	0.3077	545	0.4178	0.3761	94	0.2001	0.1702
Average			**0.5386**	**0.4918**		**0.2925**	**0.2519**		**0.4154**	**0.3979**		**0.1670**	**0.1433**
IMF	ENDE	1031	0.3357	0.2719	708	0.3155	0.2542	1727	0.2111	0.1708	1342	0.1785	0.1431
	DEEN	923	0.3463	0.2784	603	0.3676	0.3084	1471	0.2298	0.1829	1079	0.2222	0.1844
	ENES	1078	0.4220	0.3636	862	0.3916	0.3318	1822	0.2635	0.2256	1575	0.2282	0.1924
	ESEN	1302	0.3493	0.2949	889	0.3588	0.3048	2096	0.2306	0.1947	1628	0.2119	0.1800
	ENFR	1079	0.3984	0.3438	879	0.3700	0.3163	1823	0.2464	0.2117	1593	0.2157	0.1827
	FREN	1375	0.3189	0.2713	955	0.3220	0.2775	2144	0.2143	0.1805	1675	0.1953	0.1666
	ENPT	557	0.3423	0.3016	430	0.3055	0.2674	847	0.2348	0.2066	708	0.1963	0.1709
	PTEN	531	0.3626	0.3126	343	0.3899	0.3353	763	0.2644	0.2254	557	0.2541	0.2154
Average			**0.3594**	**0.3048**		**0.3526**	**0.2995**		**0.2369**	**0.1998**		**0.2128**	**0.1794**
UN	ENDE	629	0.4246	0.3709	211	0.2612	0.2195	1172	0.1900	0.1689	515	0.0877	0.0718
	DEEN	426	0.4246	0.3709	123	0.2617	0.2195	750	0.2705	0.236	266	0.1493	0.1278
	ENES	629	0.5262	0.4658	245	0.4570	0.3959	1173	0.3161	0.2779	583	0.2192	0.1852
	ESEN	630	0.5422	0.4841	245	0.4820	0.4204	1143	0.3341	0.2974	561	0.2382	0.2085
	ENFR	550	0.4004	0.3600	263	0.1912	0.1787	1025	0.2374	0.2136	611	0.0981	0.0883
	FREN	450	0.4843	0.4311	117	0.2856	0.2564	737	0.3310	0.2944	268	0.1695	0.1530
	ENIT	629	0.4744	0.4229	247	0.4011	0.3522	1172	0.2781	0.2474	591	0.1868	0.1624
	ITEN	581	0.5106	0.4630	206	0.4926	0.4466	997	0.3250	0.2939	440	0.2571	0.2318
Average			**0.4734**	**0.4211**		**0.3541**	**0.3112**		**0.2852**	**0.2537**		**0.1757**	**0.1536**

5 Discussion

Measuring translation effectiveness using bilingual dictionaries is subject to a variety of issues: correct translations which are missing; mistranslations; and, an inability to give partial credit for minor mistakes in spelling, gender, or leading articles or prepositions. Because we require an exact lexicographic match we inappropriately score some good translations as incorrect. To illustrate, here are some examples we observed when using the UN English/French dictionary: (1) 'vending machines' was given as 'distributeur automatique', not 'distributeurs automatiques' as our method produced; (2) 'urban development' was mistranslated as 'développement urbaine' (there should be no terminal 'e'); and, (3) our translation of 'tobacco products', 'produits du tabac' used 'du' instead of 'de' and thus was marked completely wrong.

Because of such issues, our method's performance is almost certainly higher than that reported in our tables and figures. To avoid this lack of sensitivity it would have been reasonable to have bilingual assessors score purported translations instead of relying only on electronic wordlists; however, we wanted to work with multiple language pairs and manual evaluation was not feasible for this study. To provide a loose upper bound on the performance attainable with imperfect dictionaries, Table 4 shows how often a 'correct' translation for an attested English term was present anywhere in the target data. Even with perfect alignments, on average only 48.9% (MWEs) and 68.8% (all words) would be marked correct.

Table 4. Limit on translation performance attainable given the bitext, by dataset.

	Pair	MWEs	All Words
FOOD	ENDE	34.3%	66.7%
	ENES	34.4%	65.5%
	ENFR	40.6%	71.4%
	ENIT	56.3%	75.0%
IFCC	ENES	51.9%	81.2%
IMF	ENDE	62.1%	70.0%
	ENES	70.2%	78.4%
	ENFR	69.1%	78.5%
	ENPT	59.5%	68.5%
UN	ENDE	27.2%	48.6%
	ENES	59.6%	72.7%
	ENFR	23.7%	52.7%
	ENIT	47.1%	65.4%
Average		**48.9%**	**68.8%**

Despite our somewhat impaired ability to correctly score translations using bilingual lexisons, we are able to report that the proposed method is demonstrably effective on terms occurring more than 10 or so times where performance of 50-70% in MRR was observed.

We briefly explored the use of Mutual Information in place of Dice scores, but found a slight drop in performance, a 2% relative decrease in MRR.

The size of aligned corpora that we were able to use was limited by available memory. The memory footprint chiefly consists of the text and 8 bytes/word to hold a word position array, and the suffix array. It is possible that algorithms for external suffix array construction could be employed, such as the DC3 algorithm by Dementiev *et al.* (2005) so that even larger corpora could be used.

We have not yet compared accuracy with results obtained from a phrase-based SMT system so we can make no claims about the relative efficacy of the two approaches; however, it is quite plausible that an iteratively trained SMT system will outperform a term similarity approach like the one we have described. An advantage in our approach is simplicity. Suffix arrays for hundreds of megabytes of text can be constructed in a couple of minutes on today's hardware and we have demonstrated that computing translations of individual phrases can be done efficiently.

6 Conclusions

We have demonstrated a new method for attaining translations using parallel data that is not intrinsically word-based but which also seamlessly projects multiword expressions across parallel texts. We investigated translation efficacy as a function of source term frequency and observed good performance (between 50% and 70%) for medium and high frequency terms. The technique is even more accurate, but this is difficult to more precisely quantify without manual review.

The major contribution of this work is describing an original method for producing translations that has been tested in multiple languages and which is effective for both isolated words and multiword expressions. We also illustrated the effects of term frequency on translation efficacy. And lastly we demonstrated a method for measuring alignment effectiveness using electronic dictionaries in place of human judgments and described some of the pitfalls that occur with this kind of evaluation.

In future work we hope to study the effect of corpus size and corpus diversity on translation effectiveness. We also hope to focus on evaluation of longer MWEs (*i.e.,* trigrams and longer) as well as consider the possibility that efficient suffix-based wildcard searches (Gusfield, 1997) may enable correct translation of non-contiguous phrases.

Acknowledgments

We are thankful for the contributions of several anonymous reviewers who offered pointers to other research and suggestions to improve the presentation of this work.

References

I. Blank. 2000. Terminology Extraction from Parallel Technical Texts. In J. Veronis (ed.) *Parallel Text Processing: Alignment and Use of Parallel Corpora*, Kluwer Academic.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263-311.

C. Callison-Burch, C. Bannard, and J. Schroeder. 2005. Scaling Phrase-Based Statistical Machine Translation to Larger Corpora and Longer Phrases. *In Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics*.

K. W. Church. 1995. One term or two?. In *Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR-95),* pp. 310-318.

I. Dagan and K. Church. 1997. Termight: Coordinating man and machine in bilingual terminology acquisition. *Machine Translation*, 12(1-2):89-107.

R. Dementiev, J. Kärkkäinen, J. Mehnert and P. Sanders. 2005. Better external memory suffix array construction. In the *Proceedings of the 7th Workshop on Algorithm Engineering and Experiments (ALENEX '05),* SIAM.

J.L. Fagan. 1998. Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods, PhD Thesis, Cornell University.

D. Gusfield. 1997. *Algorithms on Strings, Trees, and Graphs*. Cambridge University Press.

J. Kärkkäinen and P. Sanders. 2003. Simple linear work suffix array construction. In *Proc. 30th International Colloquium on Automata, Languages and Programming (ICALP '03)*. LNCS 2719, Springer, pp. 943-955.

P. Ko and S. Aluru. 2003. Space efficient linear time construction of suffix arrays. *In Combinatorial Pattern Matching (CPM 03)*. LNCS 2676, Springer, pp 203-210.

P. Koehn. 2003. Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Unpublished manuscript available online from http://people.csail.mit.edu/koehn/publications/europarl/

J. Kupiec. 1993. An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. *Proceedings of the 31st Conference of the Association for Computational Linguistics*, pp. 17-22, 1993.

J. Larsson and S. Kunihiko. 1999. Faster Suffix Sorting. Department of Computer Science, Lund University, Technical Report LU-CS-TR:99-214, Sweden.

U. Manber and G. Myers. 1991. Suffix arrays: a new method for on-line string searchers. *SIAM Journal on Computing*, 22(5):935-948.

D. S. Munteanu and D. Marcu. 2002. Processing Comparable Corpora with Bilingual Suffix Trees. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002),* Philadelphia, PA.

F. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30:417-449.

S. J. Puglisi, W. F. Smyth, A. Turpin. 2005. The Performance of Linear Time Suffix Sorting Algorithms. *Data Compression Conference (DCC'05),* pp. 358-367.

F. Smadja, K. McKeown, and V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1):1-38.

M. Yamamoto and K. W. Church. 2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, 27(1):1-30.

Y. Zhang and S. Vogel. 2005. An Efficient Phrase-to-Phrase Alignment Model for Arbitrarily Long Phrases and Large Corpora. *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT 2005)*.

G. Zipf. 1949. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, Cambridge MA.