# Bilingual Lexicon Generation Using Non-Aligned Signatures

**Daphna Shezaf**
Institute of Computer Science
Hebrew University of Jerusalem
daphna.shezaf@mail.huji.ac.il

**Ari Rappoport**
Institute of Computer Science
Hebrew University of Jerusalem
arir@cs.huji.ac.il

## Abstract

Bilingual lexicons are fundamental resources. Modern automated lexicon generation methods usually require parallel corpora, which are not available for most language pairs. Lexicons can be generated using non-parallel corpora or a pivot language, but such lexicons are noisy. We present an algorithm for generating a high quality lexicon from a noisy one, which only requires an independent corpus for each language. Our algorithm introduces *non-aligned signatures (NAS)*, a cross-lingual word context similarity score that avoids the over-constrained and inefficient nature of alignment-based methods. We use NAS to eliminate incorrect translations from the generated lexicon. We evaluate our method by improving the quality of noisy Spanish-Hebrew lexicons generated from two pivot English lexicons. Our algorithm substantially outperforms other lexicon generation methods.

## 1 Introduction

Bilingual lexicons are useful for both end users and computerized language processing tasks. They provide, for each source language word or phrase, a set of translations in the target language, and thus they are a basic component of dictionaries, which also include syntactic information, sense division, usage examples, semantic fields, usage guidelines, etc.

Traditionally, when bilingual lexicons are not compiled manually, they are extracted from parallel corpora. However, for most language pairs parallel bilingual corpora either do not exist or are at best small and unrepresentative of the general language.

Bilingual lexicons can be generated using non-parallel corpora or pivot language lexicons (see Section 2). However, such lexicons are noisy. In this paper we present a method for generating a high quality lexicon given such a noisy one. Our evaluation focuses on the pivot language case.

Pivot language approaches deal with the scarcity of bilingual data for most language pairs by relying on the availability of bilingual data for each of the languages in question with a third, *pivot*, language. In practice, this third language is often English.

A naive method for pivot-based lexicon generation goes as follows. For each source headword[1], take its translations to the pivot language using the source-to-pivot lexicon, then for each such translation take its translations to the target language using the pivot-to-target lexicon. This method yields highly noisy ('divergent') lexicons, because lexicons are generally intransitive. This intransitivity stems from polysemy in the pivot language that does not exist in the source language. For example, take French-English-Spanish. The English word *spring* is the translation of the French word *printemps*, but only in the season of year sense. Further translating *spring* into Spanish yields both the correct translation *primavera* and an incorrect one, *resorte* (the elastic object).

To cope with the issue of divergence due to lexical intransitivity, we present an algorithm for assessing the correctness of candidate translations. The algorithm is quite simple to understand and to implement and is computationally efficient. In spite of its simplicity, we are not aware of previous work applying it to our problem.

The algorithm utilizes two monolingual corpora, comparable in their domain but otherwise unrelated, in the source and target languages. It does not need a pivot language corpus. The algorithm comprises two stages: signature genera-

[1] In this paper we focus on single word head entries. Multi-word expressions form a major topic in NLP and their handling is deferred to future work.



*Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 98–107,
Uppsala, Sweden, 11-16 July 2010. 

tion and signature ranking. The *signature* of word $w$ is the set of words that co-occur with $w$ most strongly. While co-occurrence scores are used to compute signatures, signatures, unlike context vectors, do not contain the score values. For each given source headword we compute its signature and the signatures of all of its candidate translations. We present the *non-aligned signatures (NAS)* similarity score for signature and use it to rank these translations. *NAS* is based on the number of headword signature words that may be translated using the input noisy lexicon into words in the signature of a candidate translation.

We evaluate our algorithm by generating a bilingual lexicon for Hebrew and Spanish using pivot Hebrew-English and English-Spanish lexicons compiled by a professional publishing house. We show that the algorithm outperforms existing algorithms for handling divergence induced by lexical intransitivity.

## 2 Previous Work

### 2.1 Parallel Corpora

Parallel corpora are often used to infer word-oriented machine-readable bilingual lexicons. The texts are aligned to each other, at chunk- and/or word-level. Alignment is generally evaluated by consistency (source words should be translated to a small number of target words over the entire corpus) and minimal shifting (in each occurrence, the source should be aligned to a translation nearby). For a review of such methods see (Lopez, 2008). The limited availability of parallel corpora of sufficient size for most language pairs restricts the usefulness of these methods.

### 2.2 Pivot Language Without Corpora

#### 2.2.1 Inverse Consultation

Tanaka and Umemura (1994) generated a bilingual lexicon using a pivot language. They approached lexical intransitivity divergence using *Inverse Consultation (IC)*. IC examines the intersection of two pivot language sets: the set of pivot translations of a source-language word $w$, and the set of pivot translations of each target-language word that is a candidate for being a translation to $w$. IC generally requires that the intersection set contains at least two words, which are synonyms. For example, the intersection of the English translations of French *printemps* and Spanish *resorte* contains only a single word, *spring*. The intersection for a correct translation pair *printemps* and *primavera* may include two synonym words, *spring* and *springtime*. Variations of this method were proposed by (Kaji and Aizono, 1996; Bond et al., 2001; Paik et al., 2004; Ahn and Frampton, 2006).

One weakness of IC is that it relies on pivot language synonyms to identify correct translations. In the above example, if the relatively rare *springtime* had not existed or was missing from the input lexicons, IC would not have been able to discern that *primavera* is a correct translation. This may result in low recall.

#### 2.2.2 Multiple Pivot Languages

Mausam et al. (2009) used many input bilingual lexicons to create bilingual lexicons for new language pairs. They represent the multiple input lexicons in a single undirected graph, with words from all the lexicons as nodes. The input lexicons translation pairs define the edges in the graph. New translation pairs are inferred based on cycles in the graph, that is, the existence of multiple paths between two words in different languages.

In a sense, this is a generalization of the pivot language idea, where multiple pivots are used. In the example above, if both English and German are used as pivots, *printemps* and *primavera* would be accepted as correct because they are linked by both English *spring* and German *Fruehling*, while *printemps* and *resorte* are not linked by any German pivot. This multiple-pivot idea is similar to *Inverse Consultation* in that multiple pivots are required, but using multiple pivot *languages* frees it from the dependency on rich input lexicons that contain a variety of synonyms. This is replaced, however, with the problem of coming up with multiple suitable input lexicons.

#### 2.2.3 Micro-Structure of Dictionary Entries

Dictionaries published by a single publishing house tend to partition the semantic fields of headwords in the same way. Thus the first translation of some English headword in the English-Spanish and in the English-Hebrew dictionaries would correspond to the same sense of the headword, and would therefore constitute translations of each other. The applicability of this method is limited by the availability of machine-readable dictionaries produced by the same publishing house. Not surprisingly, this method has been proposed by lexicographers working in such companies (Sk-

oumalova, 2001).

### 2.3 Cross-lingual Co-occurrences in Lexicon Construction

Rapp (1999) and Fung (1998) discussed semantic similarity estimation using cross-lingual context vector alignment. Both works rely on a pre-existing large (16-20K entries), correct, one-to-one lexicon between the source and target languages, which is used to align context vectors between languages. The context vector data was extracted from comparable (monolingual but domain-related) corpora. Koehn and Knight (2002) were able to do without the initial large lexicon by limiting themselves to related languages that share a writing system, and using identically-spelled words as context words. Garera et al. (2009) and Pekar et al. (2006) suggested different methods for improving the context vectors data in each language before aligning them. Garera et al. (2009) replaced the traditional window-based co-occurrence counting with dependency-tree based counting, while Pekar et al. (2006) predicted missing co-occurrence values based on similar words in the same language. In the latter work, the one-to-one lexicon assumption was not made: when a context word had multiple equivalents, it was mapped into all of them, with the original probability equally distributed between them.

**Pivot Language.** Using cross-lingual co-occurrences to improve a lexicon generated using a pivot language was suggested by Tanaka and Iwasaki (1996). Schafer and Yarowsky (2002) created lexicons between English and a target local language (e.g. Gujarati) using a related language (e.g. Hindi) as pivot. An English pivot lexicon was used in conjunction with pivot-target cognates. Cross-lingual co-occurrences were used to remove errors, together with other cues such as edit distance and Inverse Document Frequencies (IDF) scores. It appears that this work assumed a single alignment was possible from English to the target language.

Kaji et al. (2008) used a pivot English lexicon to generate initial Japanese-Chinese and Chinese-Japanese lexicons, then used co-occurrences information, aligned using the initial lexicon, to identify correct translations. Unlike other works, which require alignments of pairs (i.e., two co-occurring words in one language translatable into two co-occurring words in the other), this method relies on alignments of 3-word cliques in each language, every pair of which frequently co-occurring. This is a relatively rare occurrence, which may explain the low recall rates of their results.

## 3 Algorithm

Our algorithm transforms a noisy lexicon into a high quality one. As explained above, in this paper we focus on noisy lexicons generated using pivot language lexicons. Other methods for obtaining an initial noisy lexicon could be used as well; their evaluation is deferred to future work.

In the setting evaluated in this paper, we first generate an initial noisy lexicon *iLex* possibly containing many translation candidates for each source headword. *iLex* is computed from two pivot-language lexicons, and is the only place in which the algorithm utilizes the pivot language. Afterwards, for each source headword, we compute its signature and the signatures of each of its translation candidates. Signature computation utilizes a monolingual corpus to discover the words that are most strongly related to the word. We now rank the candidates according to the non-aligned signatures (NAS) similarity score, which assesses the similarity between each candidate's signature and that of the headword. For each headword, we select the $t$ translations with the highest NAS scores as correct translations.

### 3.1 Input Resources

The resources required by our algorithm as evaluated in this paper are: (a) two bilingual lexicons, one from the source to the pivot language and the other from the pivot to the target language. In principle, these two pivot lexicons can be noisy, although in our evaluation we use manually compiled lexicons; (b) two monolingual corpora, one for each of the source and target languages. We have tested the method with corpora of comparable domains, but not covering the same well-defined subjects (the corpora contain news from different countries and over non-identical time periods).

### 3.2 Initial Lexicon Construction

We create an initial lexicon from the source to the target language using the pivot language: we look up each source language word $s$ in the source-pivot lexicon, and obtain the set $P_s$ of its pivot

translations. We then look up each of the members of $P_s$ in the pivot-target lexicon, and obtain a set $T_s$ of candidate target translations. *iLex* is therefore a mapping from the set of source headwords to the set of candidate target translations. Note that it is possible that not all target lexicon words appear as translation candidates. To create a target to source lexicon, we repeat the process with the directions reversed.

### 3.3 Signatures

The *signature* of a word $w$ in a language is the set of *N* words most strongly related to $w$. There are various possible ways to formalize this notion. We use a common and simple one, the words having the highest tendency to co-occur with *w* in a corpus. We count co-occurrences using a sliding fixed-length window of size *k*. We compute, for each pair of words, their Pointwise Mutual Information (PMI), that is:

$$PMI(w_1, w_2) = log\frac{Pr(w_1, w_2)}{Pr(w_1)Pr(w_2)}$$

where $Pr(w_1, w_2)$ is the co-occurrence count, and $Pr(w_i)$ is the total number of appearance of $w_i$ in the corpus (Church and Hanks, 1990). We define the signature $G(w)_{N,k}$ of *w* to be the set of *N* words with the highest PMI with *w*.

Note that a word's signature includes words in the same language. Therefore, two signatures of words in different languages cannot be directly compared; we compare them using a lexicon $L$ as explained below.

Signature is a function of $w$ parameterized by $N$ and $k$. We discuss the selection of these parameters in section 4.1.5.

### 3.4 Non-aligned Signatures (NAS) Similarity Scoring

The core strength of our method lies in the way in which we evaluate similarity between words in the source and target languages. For a lexicon *L*, a source word *s* and a target word *t*, $NAS_L(s, t)$ is defined as the number of words in the signature $G(s)_{N,k}$ of *s* that may be translated, using *L*, to words in the signature $G(t)_{N,k}$ of *t*, normalized by dividing it by *N*. Formally,

$$NAS_L(s, t) = \frac{|\{w \in G(s) | L(w) \cap G(t) \neq \emptyset\}|}{N}$$

Where $L(x)$ is the set of candidate translations of $x$ under the lexicon $L$. Since we use a single lexicon, *iLex*, throughout this work, we usually omit the *L* subscript when referring to NAS.

| Language | Sites | Tokens |
|---|---|---|
| Hebrew | haartz.co.il, ynet.co.il, nrg.co.il | 510M |
| Spanish | elpais.com, elmundo.com, abc.es | 560M |

Table 1: Hebrew corpus data.

## 4 Lexicon Generation Experiments

We tested our algorithm by generating bilingual lexicons for Hebrew and Spanish, using English as a pivot language. We chose a language pair for which basically no parallel corpora exist[2], and that do not share ancestry or writing system in a way that can provide cues for alignment.

We conducted the test twice: once creating a Hebrew-Spanish lexicon, and once creating a Spanish-Hebrew one.

### 4.1 Experimental Setup

#### 4.1.1 Corpora

The Hebrew and Spanish corpora were extracted from Israeli and Spanish newspaper websites respectively (see table 1 for details). Crawling a small number of sites allowed us to use special-tailored software to extract the textual data from the web pages, thus improving the quality of the extracted texts. Our two corpora are comparable in their domains, news and news commentary.

No kind of preprocessing was used for the Spanish corpus. For Hebrew, closed-class words that are attached to the succeeding word (e.g., 'the', 'and', 'in') were segmented using a simple unsupervised method (Dinur et al., 2009). This method compares the corpus frequencies of the non-prefixed form *x* and the prefixed form *wx*. If *x* is frequent enough, it is assumed to be the correct form, and all the occurrences of *wx* are segmented into two tokens, *w x*. This method was chosen for being simple and effective. However, the segmentation it produces is not perfect. It is context insensitive, segmenting all appearances of a token in the same way, while many *wx* forms are actually ambiguous. Even unambiguous token segmentations may fail when the non-segmented form is very frequent in the domain.

[2]Old testament corpora are for biblical Hebrew, which is very different from modern Hebrew.

| Lexicon | # headwords | BF |
|---|---|---|
| Eng-Spa | 55057 | 2.4 |
| Spa-Eng | 44349 | 2.9 |
| Eng-Heb | 48857 | 2.5 |
| Heb-Eng | 33439 | 3.7 |
| Spa-Heb | 34077 | 12.6 |
| Heb-Spa | 27591 | 14.8 |

Table 2: Number of words in lexicons, and branching factors (BF).

Hebrew orthography presents additional difficulties: there are relatively many homographs, and spelling is not quite standardized. These considerations lead us to believe that our choice of language pair is more challenging than, for example, a pair of European languages.

#### 4.1.2 Lexicons

The source of the Hebrew-English lexicon was the Babylon on-line dictionary[3]. For Spanish-English, we used the union of Babylon with the Oxford English-Spanish lexicon. Since the corpus was segmented to words using spaces, lexicon entries containing spaces were discarded.

Lexicon directionality was ignored. All translation pairs extracted for Hebrew-Spanish via English, were also reversed and added to the Spanish-Hebrew lexicon, and vice-versa. Therefore, every *L1-L2* lexicon we mention is identical to the corresponding *L2-L1* lexicon in the set of translation pairs it contains. Our lexicon is thus the 'noisiest' that can be generated using a pivot language and two source-pivot-target lexicons, but it also provides the most complete candidate set possible. Ignoring directionality is also in accordance with the *reversibility principle* of the lexicographic literature (Tomaszczyk, 1998).

Table 2 details the sizes and branching factors (BF) (the average number of translations for headword) of the input lexicons, as well as those of the generated initial noisy lexicon.

#### 4.1.3 Baseline

The performance of our method was compared to three baselines: Inverse Consultation (IC), average cosine distance, and average city block distance. The first is a completely different algorithm, and the last two are a version of our algorithm in which the NAS score is replaced by other scores.

IC (see section 2.2.1) is a corpus-less method. It ranks $t_1, t_2, ...$, the candidate translations of a source word $s$, by the size of the intersections of the sets of *pivot* translations of $t_i$ and $s$. Note that IC ranking is a partial order, as the intersection size may be the same for many candidate translations. IC is a baseline for our algorithm as a whole.

Cosine and city block distances are widely used methods for calculating distances of vectors within the same vector space. They are defined here as[4]

$$Cosine(v,u) = 1 - \frac{\sum v_i u_i}{\sqrt{\sum v_i \sum u_i}}$$

$$CityBlock(v,u) = -\sum_i |v_i - u_i|$$

In the case of context vectors, the vector indices, or keys, are words, and their values are co-occurrence based scores. We used the words in our signatures as context vector keys, and PMI scores as values. In this way, the two scores are 'plugged' into our method and serve as baselines for our NAS similarity score.

Since the context vectors are in different languages, we had to translate, or align, the baseline context vectors for the source and target words. Our initial lexicon is a many-to-many relation, so multiple alignments were possible; in fact, the number of possible alignments tends to be very large[5]. We therefore generated *M* random possible alignments, and used the average distance metric across these alignments.

#### 4.1.4 Test Sets and Gold Standard

Following other works (e.g. (Rapp, 1999)), and to simplify the experimental setup, we focused in our experiments on nouns.

A *p-q frequency range* in a corpus is the set of tokens in the places between $p$ and $q$ in the list of corpus tokens, sorted by frequency from high to low. Two types of test sets were used. The first (R1) includes all the singular, correctly segmented (in Hebrew) nouns among the 500 words in the 1001-1500 frequency range. The 1000 highest-frequency tokens were discarded, as a large number of these are utilized as auxiliary syntactic

[3] www.babylon.com.

[4] We modified the standard cosine and city block metrics so that for all measures higher values would be better.

[5] This is another advantage of our NAS score.

| | R1 | | R2 | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| NAS | **82.1%** | **100%** | **56%** | **100%** |
| Cosine | 60.7% | 100% | 28% | 100% |
| City block | 56.3% | 100% | 32% | 100% |
| IC | 55.2% | 85.7% | 52% | 88% |

Table 3: Hebrew-Spanish lexicon generation: highest-ranking translation.

| | R1 | | R2 | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| NAS | **87.6%** | **100%** | **80%** | **100%** |
| Cosine | 68% | 100% | 44% | 100% |
| City block | 69.8% | 100% | 36% | 100% |
| IC | 76.4% | 100% | 48% | 92% |

Table 4: Spanish-Hebrew Lexicon Generation: highest-ranking translation.

words. This yielded a test set of 112 Hebrew nouns and 169 Spanish nouns. The second (R2), contains 25 words for each of the two languages, obtained by randomly selecting 5 singular correctly segmented nouns from each of the 5 frequency ranges 1-1000 to 4001-5000.

For each of the test words, the correct translations were extracted from a modern professional concise printed Hebrew-Spanish-Hebrew dictionary (Prolog, 2003). This dictionary almost always provides a single Spanish translation for Hebrew headwords. Spanish headwords had 1.98 Hebrew translations on the average. In both cases this is a small number of correct translation comparing to what we might expect with other evaluation methods; therefore this evaluation amounts to a relatively high standard of correctness. Our score comparison experiments (section 5) extend the evaluation beyond this gold standard.

#### 4.1.5 Parameters

The following parameter values were used. The window size for co-occurrence counting, $k$, was 4. This value was chosen in a small pre-test. Signature size $N$ was 200 (see Section 6.1). The number of alignments $M$ for the baseline scores was 100. The number of translations selected for each headword, $t$, was set to 1 for ease of testing, but see further notes under results.

### 4.2 Results

Tables 3 and 4 summarize the results of the Hebrew-Spanish and Spanish-Hebrew lexicon generation respectively, for both the R1 and R2 test sets.

In the three co-occurrence based methods, NAS similarity, cosine distance and and city block distance, the highest ranking translation was selected. Recall is always 100% as a translation from the candidate set is always selected, and all of this set is valid. Precision is computed as the number of test words whose selected translation was one of the translations in the gold standard.

IC translations ranking is a partial order, as usually many translations are scored equally. When *all* translations have the same score, IC is effectively undecided. We calculate recall as the percentage of cases in which there was more than one score rank. A result was counted as precise if *any* of the highest-ranking translations was in the gold-standard, even if other translations were equally ranked, creating a bias in favor of IC.

In both of the Hebrew-Spanish and the Spanish-Hebrew cases, our method significantly outperformed all baselines in generating a precise lexicon on the highest-ranking translations.

All methods performed better in *R1* than in *R2*, which included also lower-frequency words, and this was more noticeable with the corpus-based methods (Hebrew-Spanish) than with IC. This suggests, not surprisingly, that the performance of corpus-based methods is related to the amount of information in the corpus.

That the results for the Spanish-Hebrew lexicon are higher may arise from the difference in the gold standard. As mentioned, Hebrew words only had one "correct" Spanish translation, while Spanish had 1.98 correct translations on the average. If we had used a more comprehensive resource to test against, the precision of the method would be higher than shown here.

In translation pairs generation, the results beyond the top-ranking pair are also of importance. Tables 5 and 6 present the accuracy of the first three translation suggestions, for the three co-occurrence based scores, calculated for the R1 test set. IC results are not included, as they are incomparable to those of the other methods: IC tends to score many candidate translations identically, and in practice, the three highest-scoring sets of translation candidates contained on average 77% of all

| | 1st | 2nd | 3rd | total |
|---|---|---|---|---|
| NAS | **82.1%** | 6.3% | 1.8% | **90.2%** |
| Cosine | 60.7% | **9.8%** | 2.7% | 73.2% |
| City block | 56.3% | 4.5% | **10.7%** | 71.4% |

Table 5: Hebrew-Spanish lexicon generation: accuracy of 3 best translations for the R1 condition. The table shows how many of the 2nd and 3rd translations are correct. Note that NAS is always a better solution, even though its numbers for 2nd and 3rd are smaller, because its accumulative percentage, shown in the last column, is higher.

| | 1st | 2nd | 3rd | total |
|---|---|---|---|---|
| NAS | **87.6%** | **77.5%** | **16%** | **163.9%** |
| Cosine | 68% | 66.3% | 10.1% | 144.4% |
| City block | 69.8% | 64.5% | **7.7%** | 142% |

Table 6: Spanish-Hebrew lexicon generation: accuracy of 3 best translations for the R1 condition. The total exceeds 100% because Spanish words had more than one correct translation. See also the caption of Table 5.

the candidates, thus necessarily yielding mostly incorrect translations. Recall was omitted from the tables as it is always 100%.

For all methods, many of the correct translations that do not rank first, rank as second or third. For both languages, NAS ranks highest for total accuracy of the three translations, with considerable advantage.

## 5 Score Comparison Experiments

Lexicon generation, as defined in our experiment, is a relatively high standard for cross-linguistic semantic distance evaluation. This is especially correct since our gold standard gives only a small set of translations. The set of possible translations in *iLex* tends to include, besides the "correct" translation of the gold standard, other translations that are suitable in certain contexts or are semantically related. For example, for one Hebrew word, *kvuza*, the gold standard translation was *grupo* (group), while our method chose *equipo* (team), which was at least as plausible given the amount of sports news in the corpus.

| | Heb-Spa | | Spa-Heb | |
|---|---|---|---|---|
| | SCE1 | SCE2 | SCE1 | SCE2 |
| **NAS** | **93.8%** | **76.2%** | **94.1%** | **83.7%** |
| Cosine | 74.1% | 57.1% | 70.7% | 63.2% |
| City block | 74.1% | 68.3% | 78,1% | 75.2% |

Table 7: Precision of score comparison experiments. The percentage of cases in which each of the scoring methods was able to successfully distinguish the *correct* (SCE1) or *possible correct* (SCE2) translation from the *random* translation.

Thus to better compare the capability of NAS to distinguish correct and incorrect translations with that of other scores, we performed two more experiments. In the first score comparison experiment (SCE1), we used the two R1 test sets, Hebrew and Spanish, from the lexicon generation test (section 4.1.4). For each word in the test set, we used our method to select between one of two translations: a *correct translation*, from the gold standard, and a *random* translation, chosen randomly among all the nouns similar in frequency to the correct translation.

The second score comparison experiment (SCE2) was designed to test the score with a more extensive test set. For each of the two languages, we randomly selected 1000 nouns, and used our method to select between a *possibly correct* translation, chosen randomly among the translations suggested in *iLex*, and a *random* translation, chosen randomly among nouns similar in frequency to the *possibly correct* translation. This test, while using a more extensive test set, is less accurate because it is not guaranteed that any of the input translations is correct.

In both SCE1 and SCE2, cosine and city block distance were used as baselines. Inverse Consultation is irrelevant here because it can only score translation pairs that appear in *iLex*.

Table 7 presents the results of the two score comparison experiments, each of them for each of the translation directions. Recall is by definition 100% and is omitted.

Again, NAS performs better than the baselines in all cases. With all scores, precision values in SCE1 are higher than in the lexicon generation experiment. This is consistent with the expectation that selection between a correct and a random, probably incorrect, translation is easier than selecting among the translations in *iLex*. The precision in SCE2 is lower than that in SCE1. This may be a result of both translations in SCE2 being

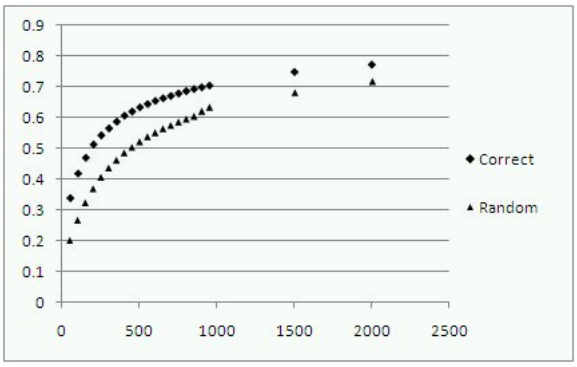


Figure 1: NAS values (not algorithm precision) for various N sizes. NAS is not sensitive to the value of N (see text).

in some cases incorrect. Yet this may also reflect a weakness of all three scores with lower-frequency words, which are represented in the 1000-word samples but not in the ones used in SCE1.

## 6 NAS Score Properties

### 6.1 Signature Size

NAS values are in the range $[0, 1]$. The values depend on *N*, the size of the signature used. With an extremely small *N*, NAS values would usually be 0, and would tend to be noisy, due to accidental inclusion of high-frequency or highly ambiguous words in the signature. As *N* approaches the size of the lexicon used for alignment, NAS values approach 1 for all word pairs.

This suggests that choosing a suitable value of *N* is critical for effectively using NAS. Yet an empirical test has shown that NAS may be useful for a wide range of *N* values: we computed NAS values for the *correct* and *random* translations used in the Hebrew-Spanish SCE1 experiment (section 5), using *N* values between 50 and 2000.

Figure 1 shows the average score values (note that these are not precision values) for the correct and random translations across that *N* range. The scores for the correct translations are consistently higher than those of the random translations, even while there is a discernible decline in the difference between them. In fact, the precision of the selection between the correct and random translation is persistent throughout the range. This suggests that while extreme *N* values should be avoided, the selection of *N* is not a major issue.

### 6.2 Dependency on Alignment Lexicon

$NAS_L$ values depend on $L$, the lexicon in use. Clearly again, in the extremes, an almost empty lexicon or a lexicon containing every possible pair of words (a Cartesian product), this score would not be useful. In the first case, it would yield 0 for every pair, and in the second, 1. However as our experiments show, it performed well with real-world examples of a noisy lexicon, with branching factors of 12.6 and 14.8 (see table 2).

### 6.3 Lemmatization

Lemmatization is the process of extracting the lemmas of words in the corpus. Our experiments show that good results can be achieved without lemmatization, at least for nouns in the pair of languages tested (aside from the simple prefix segmentation we used for Hebrew, see section 4.1.1). For other language pairs lemmatization may be needed. In general, correct lemmatization should improve results, since the signatures would consist of more meaningful information. If automatic lemmatization introduces noise, it may reduce the results' quality.

### 6.4 Alternative Models for Relatedness

Cosine and city block, as well as other related distance metrics, rely on *context vectors*. The context vector of a word *w* collects words and maps them to some score of their "relatedness" to *w*; in this case, we used PMI. NAS, in contrast, relies on the signature, the set of *N* words most related to *w*. That is, it requires a Boolean relatedness indication, rather than a numeric relatedness score. We used PMI to generate this Boolean indication, and naturally, other similar measures could be used as well. More significantly, it may be possible to use it with corpus-less sources of "relatedness", such as WordNet or search result snippets.

## 7 Conclusion

We presented a method to create a high quality bilingual lexicon given a noisy one. We focused on the case in which the noisy lexicon is created using two pivot language lexicons. Our algorithm uses two unrelated monolingual corpora. At the heart of our method is the non-aligned signatures (NAS) context similarity score, used for removing incorrect translations using cross-lingual co-occurrences.

Words in one language tend to have multiple translations in another. The common method for context similarity scoring utilizes some algebraic distance between context vectors, and requires a single alignment of context vectors in one language into the other. Finding a single correct alignment is unrealistic even when a perfectly correct lexicon is available. For example, alignment forces us to choose one correct translation for each context word, while in practice a few possible terms may be used interchangeably in the other language. In our task, moreover, the lexicon used for alignment was automatically generated from pivot language lexicons and was expected to contain errors.

NAS does not depend on finding a single correct alignment. While it measures how well the sets of words that tend to co-occur with these two words align to each other, its strength may lie in bypassing the question of *which* word in one language should be aligned to a certain context word in the other language. Therefore, unlike other scoring methods, it is not effected by incorrect alignments.

We have shown that NAS outperforms the more traditional distance metrics, which we adapted to the many-to-many scenario by amortizing across multiple alignments. Our results confirm that alignment is problematic in using co-occurrence methods across languages, at least in our settings. NAS constitutes a way to avoid this problem.

While the purpose of this work was to discern correct translations from incorrect one, it is worth noting that our method actually ranks translation correctness. This is a stronger property, which may render it useful in a wider range of scenarios.

In fact, NAS can be viewed as a general measure for word similarity between languages. It would be interesting to further investigate this observation with other sources of lexicons (e.g., obtained from parallel or comparable corpora) and for other tasks, such as cross-lingual word sense disambiguation and information retrieval.

## References


Kisuh Ahn and Matthew Frampton. 2006. Automatic generation of translation dictionaries using intermediary languages. In *EACL 2006 Workshop on Cross-Language Knowledge Induction*.

Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. Design and construction of a machine-tractable japanese-malay dictionary. In *MT Summit VIII: Machine Translation in the Information Age, Proceedings*, pages 53–58.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

Elad Dinur, Dmitry Davidov, and Ari Rappoport. 2009. Unsupervised concept discovery in hebrew using simple unsupervised word prefix segmentation for hebrew and arabic. In *EACL 2009 Workshop on Computational Approaches to Semitic Languages*.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction:from parallel corpora to non-parallel corpora. In *The Third Conference of the Association for Machine Translation in the Americas*.

Nikesh Garera, Chris Callison-Burch, and David Yarowsky. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *CoNLL*.

Hiroyuki Kaji and Toshiko Aizono. 1996. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *COLING*.

Hiroyuki Kaji, Shin'ichi Tamamura, and Dashtseren Erdenebat. 2008. Automatic construction of a japanese-chinese dictionary via english. In *LREC*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Michael Skinner, and Jeff Bilmes. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing*.

Kyonghee Paik, Satoshi Shirai, and Hiromi Nakaiwa. 2004. Automatic construction of a transfer dictionary considering directionality. In *COLING, Multilingual Linguistic Resources Workshop*.

Viktor Pekar, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. 2006. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20:247 – 266.

Prolog. 2003. *Practical Bilingual Dictionary: Spanish-Hebew/Hebrew-Spanish*. Israel.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *ACL*.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *CoNLL*.

Hana Skoumalova. 2001. Bridge dictionaries as bridges between languages. *International Journal of Corpus Linguistics*, 6:95–105.

Kumiko Tanaka and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *Conference on Computational linguistics*.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Conference on Computational Linguistics*.

Jerzy Tomaszczyk. 1998. The bilingual dictionary under review. In *ZuriLEX'86*.