Analyzing Methods for Improving Precision of Pivot Based Bilingual Dictionaries

Xabier Saralegi, Iker Manterola, Iñaki San Vicente
R&D Elhuyar Foundation
Zelai haundi 3, Osinalde Industrialdea
20170 Usurbil, Basque Country
{x.saralegi, i.manterola, i.sanvicente}@elhuyar.com

Abstract

An A-C bilingual dictionary can be inferred by merging A-B and B-C dictionaries using B as pivot. However, polysemous pivot words often produce wrong translation candidates. This paper analyzes two methods for pruning wrong candidates: one based on exploiting the structure of the source dictionaries, and the other based on distributional similarity computed from comparable corpora. As both methods depend exclusively on easily available resources, they are well suited to less resourced languages. We studied whether these two techniques complement each other given that they are based on different paradigms. We also researched combining them by looking for the best adequacy depending on various application scenarios.

1 Introduction

Nobody doubts the usefulness and multiple applications of bilingual dictionaries: as the final product in lexicography, translation, language learning, etc. or as a basic resource in several fields such as Natural Language Processing (NLP) or Information Retrieval (IR), too. Unfortunately, only major languages have many bilingual dictionaries. Furthermore, construction by hand is a very tedious job. Therefore, less resourced languages (as well as less-common language pairs) could benefit from a method to reduce the costs of constructing bilingual dictionaries. With the growth of the web, resources like Wikipedia seem to be a good option to extract new bilingual lexicon (Erdmann et al., 2008), but the reality is that a dictionary is quite different from an encyclopedia. Wiktionary[1] is a promising asset more oriented towards lexicography. However, the presence of less resourced languages in these kinds of resources is still relative -in Wikipedia, too-.

Another way to create bilingual dictionaries is by using the most widespread languages (e.g., English, Spanish, French...) as a bridge between less resourced languages, since most languages have some bilingual dictionary to/from a major language. These pivot techniques allow new bilingual dictionaries to be built automatically. However, as the next section will show, it is no small task because translation between words is not a transitive relation at all. The presence of polysemous or ambiguous words in any of the dictionaries involved may produce wrong translation pairs. Several techniques have been proposed to deal with these ambiguity cases (Tanaka and Umemura, 1994; Shirai and Yamamoto, 2001; Bond et al., 2001; Paik et al., 2004; Kaji et al., 2008; Shezaf and Rappoport, 2010). However, each technique has different performance and properties producing dictionaries of certain characteristics, such as different levels of coverage of entries and/or translations. The importance of these characteristics depends on the context of use of the dictionary. For example, a small dictionary containing the most basic vocabulary and the corresponding most frequent translations can be adequate for some IR and NLP tasks, tourism, or initial stages of language learning. Alternatively, a dictionary which maximizes the vocabulary coverage is more oriented towards advanced users or translation services.

This paper addresses the problem of pruning

[1] http://www.wiktionary.org/

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 846–856,
Edinburgh, Scotland, UK, July 27–31, 2011.

wrong translations when building bilingual dictionaries by means of pivot techniques. We aimed to come up with a method suitable for less resourced languages. We analyzed two of the approaches proposed in the literature which are not very demanding on resources: Inverse Consultation (IC) (Tanaka and Umemura, 1994) and Distributional Similarity (DS) (Kaji et al., 2008), their strong points and weaknesses, and proposed that these two paradigms be combined. For this purpose, we studied the effect the attributes of the source dictionaries have on the performance of IC and DS-based methods, as well as the characteristics of the dictionaries produced. This could allow us to predict the performance of each method just by looking at the characteristics of the source dictionaries. Finally, we tried to provide the best combination adapted to various application scenarios which can be extrapolated to other languages.

The basis of the pivot technique is dealt with in the next section, and the state of the art in pivot techniques is reviewed in the third section. After that, the analysis of the aforementioned approaches and experiments carried out for that purpose are presented, and a proposal for combining both paradigms is included. The paper ends by drawing some conclusions from the results.

2 Pivot Technique

The basic pivot-oriented construction method is based on assuming the transitive relation of the translation of a word between two languages. Thus:

if p (pivot word) is a translation of s (source word) in the A-B dictionary and t (target word) is a translation of p in the B-C dictionary, we can say that t is therefore a translation of s, or $translation_{A,B}(s) = p$ and $translation_{B,C}(p) = t \rightarrow translation_{A,C}(s) = t$

This simplification is incorrect because it does not take into account word senses. Translations correspond to certain senses of the source words. If we look at figure 1, t (case of t_1 and t_2) can be the translation of p (p_2) for a sense c (c_3) different from the sense for which p (p_2) is the equivalent of s (c_1). This can happen when p pivot word is polysemous.

It could be thought that these causalities are not frequent, and that the performance of this basic approach could be acceptable. Let us analyze a real case. We merged a Basque-English dictionary composed of 17,672 entries and 43,021 pairs with an English-Spanish one composed of 16,326 entries and 38,128 pairs, and obtained a noised Basque-Spanish dictionary comprising 14,000 entries and 104,165 pairs. 10,000 (99,844 pairs) among all the entries have more than one translation. An automatic evaluation shows that 80.32% of these ambiguous entries contain incorrect translation equivalents (80,200 pairs out of 99,844). These results show that a basic pivot-oriented method is very sensitive to the ambiguity level of the source dictionaries. The conclusion is that the transitive relation between words across languages can not be assumed, because of the large number of ambiguous entries that dictionaries actually have. A more precise statement for the transitive property in the translation process would be:

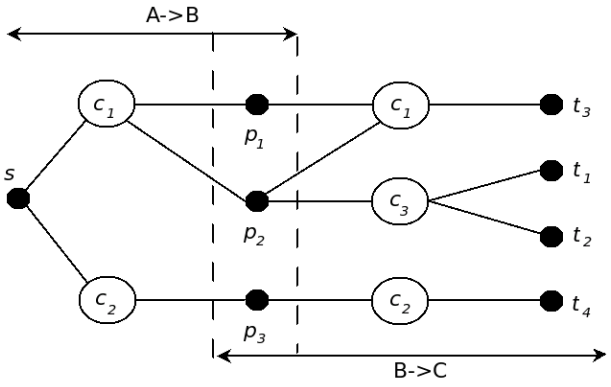

Figure 1: Ambiguity problem of the pivot technique.

if p (pivot word) is a translation of s with respect to a sense c and t is a translation of p with respect to the same sense c we can say that t is a translation of s, or $translation_{A,B}(s_{c_1}) = p$ and $translation_{B,C}(p_{c_2}) = t$ and $c_1 = c_2 \rightarrow translation_{A,C}(s) = t$

Unfortunately, most dictionaries lack comparable information about senses in their entries. So it is not possible to map entries and translation equivalents according to their corresponding senses. As an alternative, most papers try to guide this mapping according to semantic distances extracted from the dictionaries themselves or from external resources

such as corpora.

Another problem inherent in pivot-based techniques consists of missing translations. This consists of pairs of equivalents not identified in the pivot process because there is no pivot word, or else one of the equivalents is not present. We will not be dealing with this issue in this work so that we can focus on the translation ambiguity problem.

3 State of the Art

In order to reject wrong translation pairs, Tanaka et al. (1994) worked with the structure of the source dictionaries and introduced the IC method which measures the semantic distance between two words according to the number of pivot-words they share. This method was extended by using additional information from dictionaries, such as semantic classes and POS information in (Bond et al., 2001; Bond and Ogura, 2007). Sjöbergh (2005) compared full definitions in order to detect words corresponding to the same sense. However, not all the dictionaries provide this kind of information. Therefore, external knowledge needs to be used in order to guide mapping according to sense. István et al. (2009) proposed using WordNet, only for the pivot language (for English in their case), to take advantage of all the semantic information that WordNet can provide. Mausam et. al. (2009) researched the use of multiple languages as pivots, on the hypothesis that the more languages used, the more evidences will be found to find translation equivalents. They used Wiktionary for building a multilingual lexicon. Tsunakawa et al. (2008) used parallel corpora to estimate translation probabilities between possible translation pairs. Those reaching a minimum threshold are accepted as correct translations to be included in the target dictionary. However, even if this strategy achieves the best results in the terminology extraction field, it is not adequate when less resourced languages are involved because parallel corpora are very scarce.

As an alternative, (Kaji et al., 2008; Gamallo and Pichel, 2010) proposed methods to eliminate spurious translations using cross-lingual context or distributional similarity calculated from comparable corpora. In this line of work, (Shezaf and Rappoport, 2010) propose a variant of DS, and show how it outperforms the IC method. In comparison, our work focuses on analyzing the strong and weak points of each technique and aims to combine the benefits of each of them.

Other characteristics of the merged dictionaries like directionality (Paik et al., 2004) also influence the results.

4 Experimental Setup

This work focuses on adequate approaches for less resourced languages. Thus, the assumption for the experimentation is that few resources are available for both source and target languages. The resources for building the new dictionary are two basic (no definitions, no senses) bilingual dictionaries (A-B, B-C) including source (A), target (C) and a pivot language (B), as well as a comparable corpus for the source-target (A-C) language pair. We explored the IC (Tanaka and Umemura, 1994) and DS (Kaji et al., 2008; Gamallo and Pichel, 2010) approaches. In our experiments, the source and target languages are Basque and Spanish, respectively, and English is used for pivot purposes. In any case, the experiments could be conducted with any other language set, so long the required resources are available.

It must be noted that the proposed task is not a real problem because there is a Basque-Spanish dictionary already available. Resources like parallel corpora for that language pair are also available. These dictionaries and pivot language were selected in order to be able to evaluate the results automatically. During the evaluation we also used frequency information extracted from a parallel corpus, but then again, this corpus was not used during the dictionary building process, and therefore, it would not be used in a real application environment.

4.1 Resources

In order to carry out the experiments we used three dictionaries. The two dictionaries mentioned in the previous section (Basque-English $D_{eu \rightarrow en}$ and English-Spanish $D_{en \rightarrow es}$) were used to produce a new Basque-Spanish $D_{eu \rightarrow en \rightarrow es}$ dictionary. In addition, we used a Basque-Spanish $D_{eu \rightarrow es}$ dictionary for evaluation purposes. Its broad coverage is indicative of its suitability as a reference

dictionary. Table 1 shows the main characteristics of the dictionaries. We can observe that the ambiguity level of the entries (average number of translations per source word) is significant. This produces more noise in the pivot process, but it also benefits IC due to the increase in pivot words. As for the directions of source dictionaries, English is taken as target. Like Paik et al. (2004) we obtained the best coverage of pairs in that way.

Dictionary	**#entries**	**#pairs**	**ambiguity level**
$D_{eu \to en}$	17,672	43,021	2.43
$D_{en \to es}$	16,326	38,128	2.33
$D_{eu \to es}$(reference)	57,334	138,579	2.42
$D_{eu \to en \to es}$(noisy)	14,601	104,172	7.13

Table 1: Characteristics of the dictionaries.

Since we were aiming to merge two general dictionaries, the most adequate strategy was to use open domain corpora to compute DS. The domain of journalism is considered to be close to the open domain, and so we constructed a Basque-Spanish comparable corpus composed of news articles (see Table 2). The articles were gathered from the newspaper Diario Vasco (Hereinafter DV) for the Spanish part and from the Berria newspaper for the Basque part. Both publications focus on the Basque Country. In order to achieve a higher comparability degree, some constraints were applied:

- News in both languages corresponded to the same time span, 2006-2010.
- News corresponding to unrelated categories between newspapers were discarded.

Corpus	**#words**	**#docs**
Berria(eu)	40Mw	149,892
DV(es)	77Mw	306,924

Table 2: Characteristics of the comparable corpora.

In addition, as mentioned above, we extracted the frequencies of translation pairs from a Basque-Spanish parallel corpus. The corpus had 295,026 bilingual segments (4 Mw in Basque and 4.7 Mw in Spanish) from the domain of journalism.

5 Pruning Methods

IC and DS a priori suffer different weak points. IC depends on the structure of the source dictionaries. On the other hand, DS depends on a good comparable corpus and translation process. DS is measured more precisely between frequent words because context representation is richer.

The conditions for good performance of both IC and DS are analyzed below. These conditions will then be linked to the required characteristics for the initial dictionaries. In addition, we will measure how divergent the entries solved for each method are.

5.1 Inverse consultation

IC uses the structure of the D_{a-b} and D_{b-c} source dictionaries to measure the similarity of the meanings between source word and translation candidate. The description provided by Tanaka et al. (1994) is summarized as follows. To find suitable equivalents for a given entry, all target language translations of each pivot translation are looked up (e.g., $D_{b \to c}(D_{a \to b}(s))$). This way, all the "equivalence candidates" (ECs) are obtained. Then, each one is looked up in the inverse direction (following the previous example, $D_{c \to b}(t)$) to create a set of words called "selection area" (SA). The number of common elements of the same language between SA and the translations or equivalences (E) obtained in the original direction ($D_{a \to b}(s)$) is used to measure the semantic distance between entries and corresponding translations. The more matches there are, the better the candidate is. If only one inverse dictionary is consulted, the method is called "one time inverse consultation" or IC1. If n inverse dictionaries are consulted, the method is called "n time inverse consultation". As there is no significant difference in performance, we simply implemented IC1. Assuming that each element (x) of these two sets (SA, E) has a weight that is determined by the number of times it appears in the set that belongs (X), this weight is denoted as $\delta(X, x)$. In the same way, the number of common elements between SA and E is denoted as follows:

$$\delta(E, SA) = \sum_{x \in SA} \delta(E, x) \quad (1)$$

IC asks for more than one pivot word between source word s and translation candidate t. In our example:

$$\delta(D_{a\rightarrow b}(s), D_{c\rightarrow b}(t)) > 1 \quad (2)$$

In general, this condition guarantees that pivot words belong to the same sense of the source word (e.g. *iturri→tap→grifo, iturri→faucet→grifo*). Consequently, source word and target word also belong to the same sense.

Conceptually, the IC method is based on the confluence of two evidences. Let us take our dictionaries as examples. If two or more pivot words share a translation t in the $D_{es\rightarrow en}$ dictionary ($|tr(t_c, D_{es\rightarrow en}| > 1$) (e.g. *grifo→tap, grifo→faucet*) we could hypothesize that they are lexical variants belonging to a unique sense c. If an entry s includes those translations ($|tr(s_c, D_{eu\rightarrow en})| > 1$) (e.g. *iturri→tap, iturri→faucet*)) in the $D_{eu\rightarrow en}$ dictionary, we could also hypothesize the same. We can conclude that entry s and candidate t are mutual translations because the hypothesis that *"faucet"* and *"tap"* are lexical variants of the same sense c is contrasted against two evidences. This makes IC highly dependant on the number of lexical variants. Specifically, IC needs several lexical variants in the pivot language per each entry sense in both dictionaries. Assuming that wrong pairs cannot fulfill this requirement (see Formula 2) we can estimate the probabilities of the conditions for solving an ambiguous pair (s, t) where s and $t \in c$, as follows:

(a) $p(|tr(s_c, D_{a\rightarrow b})| > 1)$: Estimated by computing the average coverage of lexical variants in the pivot language for each entry in $D_{a\rightarrow b}$.

(b) $p(|tr(t_c, D_{c\rightarrow b})| > 1)$: Estimated by computing the average coverage of lexical variants in the pivot language for each entry in $D_{c\rightarrow b}$.

(c) $p(|tr(s_c, D_{a\rightarrow b}) \bigcap tr(t_c, D_{c\rightarrow b})| > 1)$: Convergence degree between translations of s and t in $D_{a\rightarrow b}$ and $D_{c\rightarrow b}$ corresponding to c.

So, in order to obtain a good performance with IC, the dictionaries used need to provide a high coverage of lexical variants per sense in the pivot language. If we assume that variants of a sense do not vary considerably between dictionaries, performance of IC in terms of recall would be estimated as follows:

$$R = p(|tr(s_c, D_{a\rightarrow b})| > 1) * p(|tr(t_c, D_{c\rightarrow b})| > 1) \quad (3)$$

We estimated the adequacy of the different dictionaries in the experimental setup according to estimations (a) and (b). Average coverage of lexical variants in the pivot language was calculated for both dictionaries. It was possible because lexical variants in the target language were grouped according to senses in both dictionaries. Only ambiguous entries were analyzed because they are the set of entries which IC must solve. In the $D_{eu\rightarrow en}$ dictionary more than 75% of senses have more than one lexical variant in the pivot language. So, $p(|tr(s_c, D_{eu\rightarrow en})| > 1) = 0.75$. In $D_{es\rightarrow en}$ this percentage (23%) is much lower. So, $p(|tr(t_c, D_{es\rightarrow en})| > 1) = 0.23$. Therefore, $D_{eu\rightarrow en}$ dictionary is more suited to the IC method than $D_{es\rightarrow en}$. As the conditions must be met in the maximum of both dictionaries, performance according to Formula 3 would be: $0.75 * 0.23 = 0.17$. This means that IC alone could solve about 17% of ambiguous entries.

5.2 Distributional Similarity

DS has been used successfully for extracting bilingual terminology from comparable corpora. The underlying idea is to identify as translation equivalents those words which show similar distributions or contexts across two corpora of different languages, assuming that this similarity is proportional to the semantic distance. In other words, establishing an equivalence between cross lingual semantic distance and translation probability. This technique can be used for pruning wrong translations produced in a pivot-based dictionary building process (Kaji et al., 2008; Gamallo and Pichel, 2010).

We used the traditional approach to compute DS (Fung, 1995; Rapp, 1999). Following the "bag-of-words" paradigm, the contexts of a word w

are represented by weighted collections of words. Those words are delimited by a window (± 5 words around w) and punctuation marks. The context words are weighted with regard to w according to the Log-likelihood ratio measure, and the context vector of w is formed. After representing word contexts in both languages, the algorithm computes for each source word the similarity between its context vector and all the context vectors corresponding to words in the target language by means of the cosine measure. To be able to compute the cross-lingual similarity, the context vectors are put in the same space by translating the vectors of the source words into the target language. This is done by using a seed bilingual dictionary. The problem is that we do not have that bilingual dictionary, since that is precisely the one we are trying to build. We propose that dictionaries extracted from our noisy dictionary ($D_{eu \rightarrow en \rightarrow es}$) be used:

- Including the unambiguous entries only
- Including unambiguous entries and selecting the most frequent candidates according to the target language corpus for ambiguous entries
- The dictionary produced by the IC1 method

The second method performed better in the tests we carried out. So, that is the method implemented for the experiments in the next section.

DS calls for several conditions in order to perform well. For solving an ambiguous translation t of a source word s, both context representations must be accurate. The higher their frequency in the comparable corpus, the richer their context representation will be. In addition to context representation, the translation quality of contexts is also a critical factor for the performance of DS. Factors can be formulated as follows if we assume big and highly comparable corpora:

(a) Precision of context representation: this can be estimated by computing the frequency of the words

(b) Precision of translation process: this can be estimated by computing the quality of the seed dictionary

6 Results

In order to evaluate the performance of each pruning method, the quality of the translations was measured according to the average precision and recall of translations per entry with respect to the reference dictionary. As we were not interested in dealing with missing translations, the reference for calculating recall was drawn up with respect to the intersection between the merged dictionary ($D_{eu \rightarrow en \rightarrow es}$) and the reference dictionary ($D_{eu \rightarrow es}$). F-score is the metric that combines both precision and recall.

We also introduced the frequency of use of both entry and pair as an aspect to take into account in the analysis of the results. It is better to deal effectively with frequent words and frequent translations than rare ones. Frequency of use of Basque words and frequency of source-target translation equivalent pairs were extracted respectively from the open domain monolingual corpus and the parallel corpus described in the previous section. Corpora were lemmatized and POS tagged in both cases in order to extract the frequency information of the lemmas.

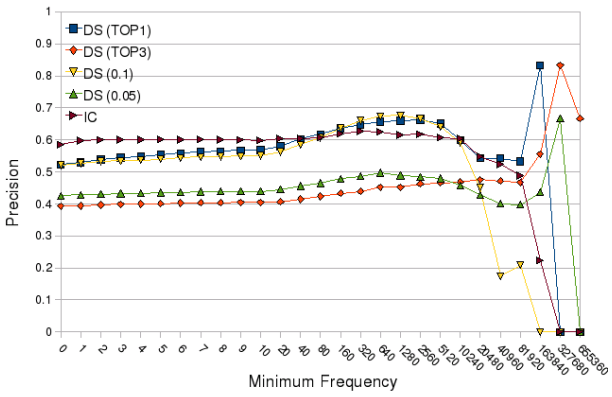

Figure 2: Precision results according to the minimum frequency of entries.

6.1 Inverse Consultation

Results show that IC precision is about 0.6 (See Figure 2). This means that many wrong pairs fulfill IC conditions. After analyzing the wrong pairs by hand, we observed that some of them corresponded to correct pairs not included in the reference dictionary. They are not included in

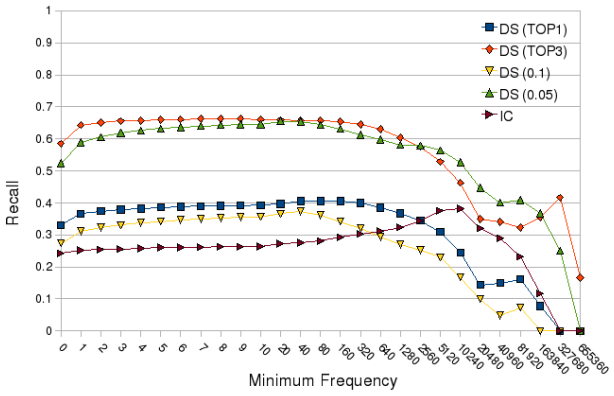

Figure 3: Recall results according to the minimum frequency of entries.

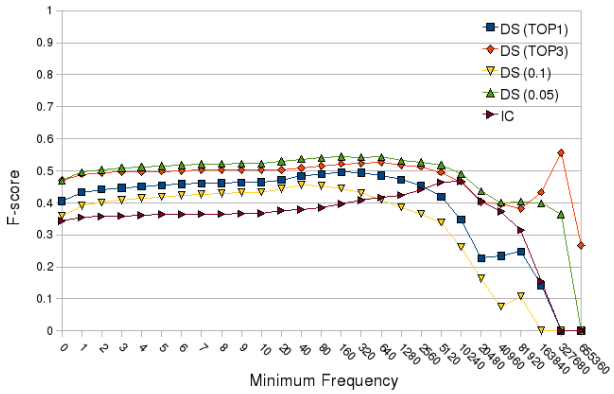

Figure 4: F-score results according to the minimum frequency of entries.

the reference because not all synonyms -or lexical variants- are included in it, only the most common ones. This is an inherent problem in automatic evaluation, and affects all the experiments presented throughout section 6 equally. Other wrong pairs comprise translation equivalents which have the same stem but different gramatical categories (e.g., 'aldakuntza' (noun) *(change, shift)* $\rightarrow$ *'cambiar'* (verb) *(to change, to shift)*). These wrong cases could be filtered if POS information would be available in the source dictionaries.

Precision is slightly better when dealing with frequent words, a maximum of 0.62 is reached when minimum frequency is between 150 and 2,000. Precision starts to decline significantly when dealing

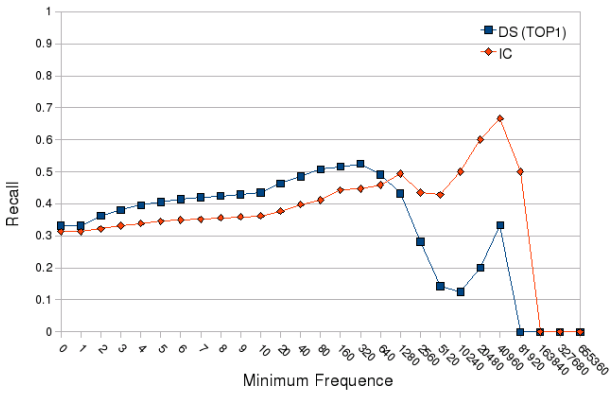

Figure 5: Recall results according to the minimum frequency of translation pairs.

with those entries over a minimum frequency of 10,000. However, only very few entries (234) reach that minimum frequency.

Recall is about 0.2 (See Figure 3), close to the estimation computed in section 5.1. It presents a more marked variability according to the frequency of entries, improving the performance as the frecuency increases. This could be due to the fact that frequent entries tend to have more translation variants (See Table 3). The fact that there are too many candidates to solve would explain why the recall starts to decline when dealing with very frequent entries.

Global performance according to F-score reflects the variability depending on frequency (See Figure 4).

Recall according to frequency of pairs provides information about whether IC selects rare translations or the most probable ones (See Figure 5). It must be noted that this recall is calculated with respect to the translation pairs of the merged dictionary $D_{eu \rightarrow en \rightarrow es}$ which appear in the parallel corpus (see section 4.1). Results (See Figure 5) show that IC deals much better with frequent translation pairs. However, recall for pairs whose frequency is higher than 100 only reaches 0.5. Even if the maximum recall is achieved for pairs whose frequency is above 40,000, it is not significant because they suppose a minimum number (3 pairs). In short, we can conclude that IC often does not find the most probable translation

(e.g. 'usain'→*'olor' (smell)*, 'zulo'→*'agujero' (hole)*,...).

6.2 Distributional Similarity

DS provides an idea of semantic distance. However, in order to determine whether a candidate is a correct translation, a minimum threshold must be established. It is very difficult to establish a threshold manually because its performance depends on the characteristics of the corpora and the seed dictionaries. The threshold can be applied at a global level, by establishing a numeric threshold for all candidates, or at local level by selecting certain top ranked candidates for each entry. The dictionary created by IC or unambiguous pairs can be used as a reference for tuning the threshold in a robust way with respect to the evaluation score such as F-score. In our experiments, thresholds estimated against the dictionary created by IC are very close to those calculated with respect to the whole reference dictionary (see Figure 6).

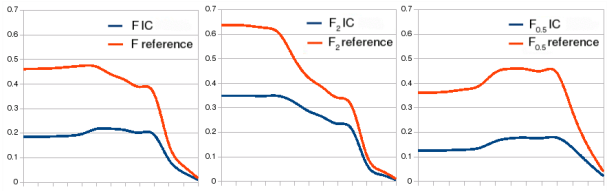

Figure 6: Threshold parameter tuning comparison for different F_n scores. Tuning against dictionary created by IC vs. Reference dictionary.

There is not much variation in performance between local and global thresholds. Precision increases from 0.4 to 0.5 depending on the strictness level of the threshold (See Figure 2), the stricter the better. In all cases, precision is slightly better when dealing with frequent words (frequency $>$ 20). This improvement is more marked with the strictest thresholds (TOP_1, 0.1). However, if global thresholds are used, performance starts to decline significantly when dealing with words whose frequency is above 1,000. So, it seems that local thresholds (TOP_3) perform more consistendly with respect to the high frequencies of entries.

Recall (See Figure 3) goes from 0.5 to 0.7 depending on the strictness level of the threshold. It starts declining when frequency is above 50 depending on the type of threshold. In this case, global thresholds seem to perform better because the most frequent entries are handled better. These entries tend to have many translations. Therefore thresholds based on top ranks are too rigid.

There is no significant difference between global and local thresholds in terms of F-Score (See Figure 4). Each threshold type is more stable in precision or recall. So the F-Score is similar for both. Variability of F-Score according to frequency is lower than in precision and recall. As performance peaks on both measures at different points of frequency, the variability is mitigated when measures are combined by F-Score.

We have plotted the recall according to the frequency of pairs calculated from a parallel corpus in order to analyze the performance of DS when dealing with frequent translation pairs (See Figure 5). The performance decreases when dealing with pairs whose frequency is higher than 100. This means that DSs performance is worse when dealing with the most common translation pairs. So it is clear that it is very difficult to represent the contexts of very frequent words correctly.

The results show that DS rankings are worse when dealing with some words above a certain frequency threshold (e.g. 'on' *'good'*, 'berriz' *'again'*, 'buru' *'head'*, 'orain' *'now'*...). Although context representation of frequent words is based on many evidences, high polysemy level related to high frequency leads to a poorer representation. Alternatively we found that some of those frequent words are not very polysemous. Those words do not have strong collocates, that is, they tend to appear freely in contexts, which also leads to poor representation. This low quality representation hampers an accurate computation of semantic distance.

6.3 Comparison between IC and DS

As for average precision, IC provides better results than DS if all entries are taken into account. However, DS tips the scales in its favor if only entries with frequencies above 50 are considered and strict thresholds are used (TOP_1, 0.1).

DS clearly outperforms IC in terms of average recall of translations. Even if strict thresholds are used, DS outperforms IC for all entries whose

frequency is lower than 640.

If average precision and recall are evaluated together by means of F-score, DS outperforms IC (Figure 4). Only when dealing with very frequent entries (frequency $> 8,000$) is ICs performance close to DSs, but these entries make up a very small group (234 entries).

In order to compare the recall with respect to the frequency of translation pairs under the same conditions, we have to select a threshold that provides a similar precision to IC. TOP_1 is the most similar one (see figure 2). As Figure 5 shows, again DS is better than IC. Even if IC's recall clearly surpasses DS's when dealing with frequent translation pairs (frequency $> 2,560$), it only represents a minimal number of pairs (39).

6.4 Combining IC and DS according to different scenarios

In order to see how the methods can complement each other, we calculated the performance for solving ambiguous entries obtained by combining the results of both methods using various alternatives:

- Union: $IC \cup DS$: Pairs obtained by both methods are merged. Duplicated pairs are cleaned.
- Lineal combination (Lcomb): $IC * k + DS * (1 - k)$. Each method provides a value representing the translation score. For IC that value is the number of pivot words (see Formula 1), and the context similarity score in the case of DS. Those values are linearly combined and applied over the noised dictionary.

As mentioned in the first section, one of the goals of the paper was to analyze which method and which combination was best depending on the use case. We have selected some measures which are a good indicator of good performance for different use cases:

- $AvgF$: Average F-score per entry.
- $wAvgF$: Average F-score per entry weighted by the frequency of the entry. Higher frequency increases the weight.
- $AvgF_2$: Average F-score per entry where recall is weighted higher.
- $AvgF_{0.5}$: Average F-score per entry where precision is weighted higher.

For the use cases presented in section 1, some measures will provide richer information than others. On the one hand, if we aim to build small, accurate dictionaries, $AvgF_{0.5}$ would be a better indicator since it attaches more importance to high precision. In addition, if we want the dictionaries to cover the most common entries (e.g., in a basic dictionary for language learners) it is also interesting to look at $wAvgF$ values because greater value is given to finding translations for the most frequent words. On the other hand, if our objective is to build big dictionaries with a high recall, it would be better to look at $AvgF_2$ measure which attaches importance to recall.

Method	$AvgF$	$wAvgF$	$AvgF_2$	$AvgF_{0.5}$
IC	0.34	0.27	0.27	0.46
DS	0.47	0.44	0.64	0.46
Union	**0.52**	**0.49**	0.65	0.49
Lcomb	**0.52**	**0.49**	**0.67**	**0.52**

Table 3: Performance results of methods for ambiguous entries according to different measures.

Table 3 shows the results for the different combinations. The parameters of all methods are optimized for each metric (as explained in section 6.2, see figure 6). In all cases, the combinations surpass the results of both methods separately. There is a reasonable improvement over DS (10.6% for $AvgF$), and an even more startling one over IC (52.9% for $AvgF$). IC only gets anywhere near the other methods when precision is given priority ($AvgF_{0.5}$). There is no significant difference in terms of performance between the two combinations, although Lcomb is slightly better. $wAvgF$ measure is stricter than the others since it takes frequency of entries into account. This is emphasised more in the case of IC where results decrease notably compared with $AvgF$.

7 Conclusions

This paper has analyzed IC and DS, for the task of pruning wrong translations from bilingual dictionaries built by means of pivot techniques. After analyzing their strong and weak points we have showed that IC requires high ambiguity level dictionaries with several lexical variants per entry sense. With an average ambiguity close to 2 translation candidates DS obtains better results. IC is a high precision method, but contrary to our expectations, it seems that it is not much more precise than DS. In addition, DS offers much better recall of translations and entries. As a result, DS performs the best if both precision and recall are taken into account by F-score.

Both methods prune most probable translations for a significant number of frequent entries. DS encounters a problem when dealing with very frequent words due to the difficulty in representing their context. The main reason behind this is the high polysemy level of those words.

Our initial beliefs were that the translations found by each method would diverge to a certain extent. The results obtained when combining the two methods show that although the performance does not increase as much as expected (10.6% improvement over DS), there is in fact some divergence. As for the different use cases proposed, combinations offer the best performance in all cases. IC is indeed the poorer method, although it presents competitive results when precision is given priority.

Future experiments include contrasting these results with other dictionaries and language pairs.

8 Aknowledgments

This work has been partially founded by the Industry Department of the Basque Government under grants IE09-262 (Berbatek project) and SA-2010/00245 (Pibolex+ project).

References

Francis Bond and Kentaro Ogura. 2007. Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary. Language Resources and Evaluation, 42(2):127–136.

Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. Design and construction of a machine-tractable Japanese-Malay dictionary. Proceedings of ASIALEX, SEOUL, 2001(2001):200–205.

Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2008. An approach for extracting bilingual terminology from wikipedia. In Proceedings of the 13th international conference on Database systems for advanced applications, DASFAA'08, pages 380–392, Berlin, Heidelberg. Springer-Verlag. ACM ID: 1802552.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In David Yarovsky and Kenneth Church, editors, Proceedings of the Third Workshop on Very Large Corpora, pages 173–183, Somerset, New Jersey. Association for Computational Linguistics.

Pablo Gamallo and José Pichel. 2010. Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010. Proceedings, volume 6008 of Lecture Notes in Computer Science, pages 473–483. Springer.

Varga István and Yokoyama Shoichi. 2009. Bilingual dictionary generation for low-resourced language pairs. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2, EMNLP '09, pages 862–870, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1699625.

Hiroyuki Kaji, Shin'ichi Tamamura, and Dashtseren Erdenebat. 2008. Automatic construction of a Japanese-Chinese dictionary via English. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

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 Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1, ACL '09, page 262270, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1687917.

Kyonghee Paik, Satoshi Shirai, and Hiromi Nakaiwa. 2004. Automatic construction of a transfer dictionary considering directionality. In Proceedings of the Workshop on Multilingual Linguistic Ressources, MLR '04, pages 31–38, Stroudsburg, PA, USA.

Association for Computational Linguistics. ACM ID: 1706243.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics, pages 519–526, College Park, USA. ACL.

Daphna Shezaf and Ari Rappoport. 2010. Bilingual lexicon generation using non-aligned signatures. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, page 98107, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1858692.

S. Shirai and K. Yamamoto. 2001. Linking english words in two bilingual dictionaries to generate another language pair dictionary. In Proceedings of ICCPOL, pages 174–179.

J. Sjöbergh. 2005. Creating a free digital Japanese-Swedish lexicon. In Proceedings of PACLING 2005.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In Proceedings of the 16th International Conference on Computational Linguistics (COLING'94), pages 297–303.

Takashi Tsunakawa, Naoaki Okazaki, and Jun'ichi Tsujii. 2008. Building bilingual lexicons using lexical translation probabilities via pivot languages. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).