Robust Measurement and Comparison of Context Similarity for Finding Translation Pairs

Daniel Andrade[†], **Tetsuya Nasukawa**[‡], **Jun'ichi Tsujii**[†]
[†]Department of Computer Science, University of Tokyo
{daniel.andrade, tsujii}@is.s.u-tokyo.ac.jp
[‡]IBM Research - Tokyo
nasukawa@jp.ibm.com

Abstract

In cross-language information retrieval it is often important to align words that are similar in meaning in two corpora written in different languages. Previous research shows that using context similarity to align words is helpful when no dictionary entry is available. We suggest a new method which selects a subset of words (pivot words) associated with a query and then matches these words across languages. To detect word associations, we demonstrate that a new Bayesian method for estimating Point-wise Mutual Information provides improved accuracy. In the second step, matching is done in a novel way that calculates the chance of an accidental overlap of pivot words using the hypergeometric distribution. We implemented a wide variety of previously suggested methods. Testing in two conditions, a small comparable corpora pair and a large but unrelated corpora pair, both written in disparate languages, we show that our approach consistently outperforms the other systems.

1 Introduction

Translating domain-specific, technical terms from one language to another can be challenging because they are often not listed in a general dictionary. The problem is exemplified in cross-lingual information retrieval (Chiao and Zweigenbaum, 2002) restricted to a certain domain. In this case, the user might enter only a few technical terms. However, jargons that appear frequently in the data set but not in general dictionaries, impair the usefulness of such systems. Therefore, various means to extract translation pairs automatically have been proposed. They use different clues, mainly

- Spelling distance or transliterations, which are useful to identify loan words (Koehn and Knight, 2002).
- Context similarity, helpful since two words with identical meaning are often used in similar contexts across languages (Rapp, 1999).

The first type of information is quite specific; it can only be helpful in a few cases, and can thereby engender high-precision systems with low recall, as described for example in (Koehn and Knight, 2002). The latter is more general. It holds for most words including loan words. Usually the context of a word is defined by the words which occur around it (bag-of-words model).

Let us briefly recall the main idea for using context similarity to find translation pairs. First, the degree of association between the query word and all content words is measured with respect to the corpus at hand. The same is done for every possible translation candidate in the target corpus. This way, we can create a feature vector for the query and all its possible translation candidates. We can assume that, for some content words, we have valid translations in a general dictionary, which enables us to compare the vectors across languages. We will designate these content words as pivot words. The query and its translation candidates are then compared using their feature vectors, where each dimension in the feature vector contains the degree of association to

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 19–27,
Beijing, August 2010

one pivot word. We define the degree of association, as a measurement for finding words that co-occur, or which do not co-occur, more often than we would expect by pure chance.[1]

We argue that common ways for comparing similarity vectors across different corpora perform worse because they assume that degree of associations are very similar across languages and can be compared without much preprocessing. We therefore suggest a new robust method including two steps. Given a query word, in the first step we determine the set of pivots that are all positively associated with statistical significance. In the second step, we compare this set of pivots with the set of pivots extracted for a possible translation candidate. For extracting positively associated pivots, we suggest using a new Bayesian method for estimating the critical Pointwise Mutual Information (PMI) value. In the second step, we use a novel measure to compare the sets of extracted pivot words which is based on an estimation of the probability that pivot words overlap by pure chance. Our approach engenders statistically significant improved accuracy for aligning translation pairs, when compared to a variety of previously suggested methods. We confirmed our findings using two very different pairs of comparable corpora for Japanese and English.

In the next section, we review previous related work. In Section 3 we explain our method in detail, and argue that it overcomes subtle weaknesses of several previous efforts. In Section 4, we show with a series of cross-lingual experiments that our method, in some settings, can lead to considerable improvement in accuracy. Subsequently in Section 4.2, we analyze our method in contrast to the baseline by giving two examples. We summarize our findings in Section 5.

2 Related Work

Extracting context similarity for nouns and then matching them across languages to find translation pairs was pioneered in (Rapp, 1999) and (Fung, 1998). The work in (Chiao and Zweigenbaum, 2002), which can be regarded as a variation of (Fung, 1998), uses tf.idf, but suggests to normalize the term frequency by the maximum number of co-occurrences of two words in the corpus. All this work is closely related to our work because they solely consider context similarity, whereas context is defined using a word window. The work in (Rapp, 1999; Fung, 1998; Chiao and Zweigenbaum, 2002) will form the baselines for our experiments in Section 4.[2] This baseline is also similar to the baseline in (Gaussier et al., 2004), which showed that it can be difficult to beat such a feature vector approach.

In principle our method is not restricted to how context is defined; we could also use, for example, modifiers and head words, as in (Garera et al., 2009). Although, we found in a preliminary experiment that using a dependency parser to differentiate between modifiers and head words like in (Garera et al., 2009), instead of a bag-of-words model, in our setting, actually decreased accuracy due to the narrow dependency window. However, our method could be combined with a back-translation step, which is expected to improve translation quality as in (Haghighi et al., 2008), which performs indirectly a back-translation by matching *all* nouns mutually exclusive across corpora. Notably, there also exist promising approaches which use both types of information, spelling distance, and context similarity in a joint framework, see (Haghighi et al., 2008), or (Déjean et al., 2002) which include knowledge of a thesaurus. In our work here, we concentrate on the use of degrees of association as an effective means to extract word translations.

In this application, to measure association robustly, often the Log-Likelihood Ratio (LLR) measurement is suggested (Rapp, 1999; Morin et al., 2007; Chiao and Zweigenbaum, 2002). The occurrence of a word in a document is modeled as a binary random variable. The LLR measurement measures stochastic dependency between

[1] For example "car" and "tire" are expected to have a high (positive) degree of association, and "car" and "apple" is expected to have a high (negative) degree of association.

[2] Notable differences are that we neglected word order, in contrast to (Rapp, 1999), as it is little useful to compare it between Japanese and English. Furthermore in contrast to (Fung, 1998) we use only one translation in the dictionary, which we select by comparing the relative frequencies. We also made a second run of the experiments where we manually selected the correct translations for the first half of the most frequent pivots – Results did not change significantly.

two such random variables (Dunning, 1993), and is known to be equal to Mutual Information that is linearly scaled by the size of the corpus (Moore, 2004). This means it is a measure for how much the occurrence of word A makes the occurrence of word B more likely, which we term positive association, and how much the absence of word A makes the occurrence of word B more likely, which we term negative association. However, our experiments show that only positive association is beneficial for aligning words cross-lingually. In fact, LLR can still be used for extracting positive associations by filtering in a pre-processing step words with possibly negative associations (Moore, 2005). Nevertheless a problem which cannot be easily remedied is that confidence estimates using LLR are unreliable for small sample sizes (Moore, 2004). We suggest a more principled approach that measures from the start only how much the occurrence of word A makes the occurrence of word B more likely, which is designated as Robust PMI.

Another point that is common to (Rapp, 1999; Morin et al., 2007; Chiao and Zweigenbaum, 2002; Garera et al., 2009; Gaussier et al., 2004) is that word association is compared in a fine-grained way, i.e. they compare the degree of association[3] with every pivot word, even when it is low or exceptionally high. They suggest as a comparison measurement Jaccard similarity, Cosine similarity, and the L1 (Manhattan) distance.

3 Our Approach

We presume that rather than similarity between *degree (strength of)* of associations, the *existence* of common word associations is a more reliable measure for word similarity because the degrees of association are difficult to compare for the following reasons:

- **Small differences in the degree of association are not statistically significant**
  Taking, for example, two sample sets from the same corpus, we will in general measure different degrees of association.
- **Differences in sub-domains / sub-topics**
  Corpora sharing the same topic can still differ in sub-topics.
- **Differences in style or language**
  Differences in word usage. [4]

Other information that is used in vector approaches such as that in (Rapp, 1999) is negative association, although negative association is less informative than positive. Therefore, if it is used at all, it should be assigned a much smaller weight.

Our approach caters to these points, by first deciding whether a pivot word is positively associated (with statistical significance) or whether it is not, and then uses solely this information for finding translation pairs in comparable corpora. It is divisible into two steps. In the first, we use a Bayesian estimated Pointwise Mutual Information (PMI) measurement to find the pivots that are positively associated with a certain word with high confidence. In the second step, we compare two words using their associated pivots as features. The similarity of feature sets is calculated using pointwise entropy. The words for which feature sets have high similarity are assumed to be related in meaning.

3.1 Extracting positively associated words – Feature Sets

To measure the degree of positive association between two words x and y, we suggest the use of information about how much the occurrence of word x makes the occurrence of word y more likely. We express this using Pointwise Mutual Information (PMI), which is defined as follows:

$$PMI(x,y) = \log \frac{p(x,y)}{p(x) \cdot p(y)} = \log \frac{p(x|y)}{p(x)} \,.$$

Therein, $p(x)$ is the probability that word x occurs in a document; $p(y)$ is defined analogously. Furthermore, $p(x,y)$ is the probability that both

[3] To clarify terminology, where possible, we will try to distinguish between *association* and *degree of association*. For example word "car" has the *association* "tire", whereas the *degree of association* with "tire" is a continuous number, like 5.6.

[4] For example, "stop" is not the only word to describe the fact that a car halted.

words occur in the same document. A positive association is given if $p(x|y) > p(x)$. In related works that use the PMI (Morin et al., 2007), these probabilities are simply estimated using relative frequencies, as

$$PMI(x,y) = \log \frac{\frac{f(x,y)}{n}}{\frac{f(x)}{n}\frac{f(y)}{n}},$$

where $f(x)$, $f(y)$ is the document frequency of word x and word y, and $f(x,y)$ is the co-occurrence frequency; n is the number of documents. However, using relative frequencies to estimate these probabilities can, for low-frequency words, produce unreliable estimates for PMI (Manning and Schütze, 2002). It is therefore necessary to determine the uncertainty of PMI estimates. The idea of defining confidence intervals over PMI values is not new (Johnson, 2001); however, the problem is that exact calculation is very computationally expensive if the number of documents is large, in which case one can approximate the binomial approximation for example with a Gaussian, which is, however only justified if n is large *and* p, the probability of an occurrence, is not close to zero (Wilcox, 2009). We suggest to define a beta distribution over each probability of the binary events that word x occurs, i.e. $[x]$, and analogously $[x|y]$. It was shown in (Ross, 2003) that a Bayesian estimate for Bernoulli trials using the beta distribution delivers good credibility intervals[5], importantly, when sample sizes are small, or when occurrence probabilities are close to 0. Therefore, we assume that

$$p(x|y) \sim beta(\alpha'_{x|y}, \beta'_{x|y}), p(x) \sim beta(\alpha'_x, \beta'_x)$$

where the parameters for the two beta distributions are set to

$$\alpha'_{x|y} = f(x,y) + \alpha_{x|y},$$
$$\beta'_{x|y} = f(y) - f(x,y) + \beta_{x|y}, \text{and}$$
$$\alpha'_x = f(x) + \alpha_x,\ \beta'_x = n - f(x) + \beta_x.$$

Prior information related to $p(x)$ and the conditional probability $p(x|y)$ can be incorporated by setting the hyper-parameters of the beta-distribtutions.[6] These can, for example, be learned from another unrelated corpora pair and then weighted appropriately by setting $\alpha + \beta$. For our experiments, we use no information beyond the given corpora pair; the conditional priors are therefore set equal to the prior for $p(x)$. Even if we do not know which word x is, we have a notion about $p(x)$ because Zipf's law indicates to us that we should expect it to be small. A crude estimation is therefore the mean word occurrence probability in our corpus as

$$\gamma = \frac{1}{|\text{all words}|} \sum_{x \in \{\text{all words}\}} \frac{f(x)}{n}.$$

We give this estimate a total weight of one observation. That is, we set

$$\alpha = \gamma, \beta = 1 - \gamma.$$

From a practical perspective, this can be interpreted as a smoothing when sample sizes are small, which is often the case for $p(x|y)$. Because we assume that $p(x|y)$ and $p(x)$ are random variables, PMI is consequently also a random variable that is distributed according to a beta distribution ratio.[7] For our experiments, we apply a general sampling strategy. We sample $p(x|y)$ and $p(x)$ independently and then calculate the ratio of times $PMI > 0$ to determine $P(PMI > 0)$.[8] We will refer to this method as Robust PMI (RPMI).

Finally we can calculate, for any word x, the set of pivot words which have most likely a positive association with word x. We require that this set be statistically significant: the probability of one or more words being not a positive association is smaller than a certain p-value.[9]

[5] In the Bayesian notation we refer here to credibility intervals instead of confidence intervals.

[6] The hyper-parameters α and β, can be intuitively interpreted in terms of document frequency. For example α_x is the number of times we belief the word x occurs, and β_x the number of times we belief that x does not occur in a corpus. Analogously $\alpha_{x|y}$ and $\beta_{x|y}$ can be interpreted with respect to the subset of the corpus where the word y occurs, instead of the whole corpus. Note however, that α and β do not necessarily have to be integers.

[7] The resulting distribution for the general case of a beta distribution ratio was derived in (Pham-Gia, 2000). Unfortunately, it involves the calculation of a Gauss hyper-geometric function that is computationally expensive for large n.

[8] For experiments, we used $100,000$ samples for each estimate of $P(PMI > 0)$.

[9] We set, for all of our experiments, the p-value to 0.01.

As an alternative for determining the probability of a positive association using $P(PMI > 0)$, we calculate LLR and assume that approximately $LLR \sim \chi^2$ with one degree of freedom (Dunning, 1993). Furthermore, to ensure that only positive association counts, we set the probability to zero if $p(x, y) < p(x) \cdot p(y)$, where the probabilities are estimated using relative frequencies (Moore, 2005). We refer to this as LLR(P); lacking this correction, it is LLR.

3.2 Comparing Word Feature Sets Across Corpora

So far, we have explained a robust means to extract the pivot words that have a positive association with the query. The next task is to find a sensible way to use these pivots to compare the query with candidates from the target corpus. A simple means to match a candidate with a query is to see how many pivots they have in common, i.e. using the matching coefficient (Manning and Schütze, 2002) to score candidates. This similarity measure produces a reasonable result, as we will show in the experiment section; however, in our error analysis, we found out that this gives a bias to candidates with higher frequencies, which is explainable as follows. Assuming that a word A has a fixed number of pivots that are positively associated, then depending on the sample size—the document frequency in the corpus—not all of these are statistically significant. Therefore, not all true positive associations are included in the feature set to avoid possible noise. If the document frequency increases, then we can extract more statistically significant positive associations and the cardinality of the feature set increases. This consequently increases the likelihood of having more pivots that overlap with pivots from the query's feature set. For example, imagine two candidate words A and B, for which feature sets of both include the feature set of the query, i.e. a complete match, however A's feature set is much larger than B's feature set. In this case, the information conveyed by having a complete match with the query word's feature set is lower in the case of A's feature set than in case of B's feature set. Therefore, we suggest its use as a basis of our similarity measure, the degree of pointwise entropy of having an estimate of m matches, as

$$\text{Information}(m, q, c) = -\log(P(matches = m)).$$

Therein, $P(matches = m)$ is the likelihood that a candidate word with c pivots has m matches with the query word, which has q pivots. Letting w be the total number of pivot words, we can then calculate that the probability that the candidate with c pivots was selected by chance

$$P(matches = m) = \frac{\binom{q}{m} \cdot \binom{w-q}{c-m}}{\binom{w}{c}}.$$

Note that this probability equals a hypergeometric distribution.[10] The smaller $P(matches = m)$ is, the less likely it is that we obtain m matches by pure chance. In other words, if $P(matches = m)$ is very small, m matches are more than we would expect to occur by pure chance.[11]

Alternatively, in our experiments, we also consider standard similarity measurements (Manning and Schütze, 2002) such as the Tanimoto coefficient, which also lowers the score of candidates that have larger feature sets.

4 Experiments

In our experiments, we specifically examine translating nouns, mostly technical terms, which occur in complaints about cars collected by the Japanese Ministry of Land, Infrastructure, Transport and Tourism (MLIT)[12], and in complaints about cars collected by the USA National Highway Traffic Safety Administration (NHTSA)[13]. We create for each data collection a corpus for which a document corresponds to one car customer reporting a certain problem in free text. The complaints are, in general, only a few sentences long.

[10] $\binom{q}{m}$ is the number of possible combinations of pivots which the candidate has in common with the query. Therefore, $\binom{q}{m} \cdot \binom{w-q}{c-m}$ is the number of possible different feature sets that the candidate can have such that it shares m common pivots with the query. Furthermore, $\binom{w}{c}$ is the total number of possible feature sets the candidate can have.

[11] The discussion is simplified here. It can also be that $P(matches = m)$ is very small, if there are less occurrences of m that we would expect to occur by pure chance. However, this case can be easily identified by looking at the gradient of $P(matches = m)$.

[12] http://www.mlit.go.jp/jidosha/carinf/rcl/defects.html

[13] http://www-odi.nhtsa.dot.gov/downloads/index.cfm

To verify whether our results can be generalized over other pairs of comparable corpora, we additionally made experiments using two corpora extracted from articles of Mainichi Shinbun, a Japanese newspaper, in 1995 and English articles from Reuters in 1997. There are two notable differences between those two pairs of corpora: the content is much less comparable, Mainichi reports more national news than world news, and secondly, Mainichi and Reuters corpora are much larger than MLIT/NHTSA.[14]

For both corpora pairs, we extracted a gold-standard semi-automatically by looking at Japanese nouns and their translations with document frequency of at least 50 for MLIT/NHTSA, and 100 for Mainichi/Reuters. As a dictionary we used the Japanese-English dictionary JMDic[15]. In general, we preferred domain-specific terms over very general terms, i.e. for example for MLIT/NHTSA the noun 噴射 "injection" was preferred over 取り付け "installation". We extracted 100 noun pairs for MLIT/NHTSA and Mainichi/Reuters, each. Each Japanese noun which is listed in the gold-standard forms a query which is input into our system. The resulting ranking of the translation candidates is automatically evaluated using the gold-standard. Therefore, synonyms that are not listed in the gold standard are not recognized, engendering a conservative estimation of the translation accuracy. Because all methods return a ranked list of translation candidates, the accuracy is measured using the rank of the translation listed in the gold-standard.[16] The Japanese corpora are preprocessed with MeCab (Kudo et al., 2004); the English corpora with Stepp Tagger (Tsuruoka et al., 2005) and Lemmatizer (Okazaki et al., 2008). As a dictionary we use the Japanese-English dictionary JMDic[17]. In line with related work (Gaussier et al., 2004), we remove a word pair (Japanese noun s, English noun t) from the dictionary, if s occurs in the gold-standard. Afterwards we define the pivot words by consulting the remaining dictionary.

[14]MLIT/MLIT has each 20,000 documents. Mainichi/Reuters corpora 75,935 and 148,043 documents, respectively.

[15]http://www.csse.monash.edu.au/ jwb/edict_doc.html

[16]In cases for which there are several translations listed for one word, the rank of the first is used.

[17]http://www.csse.monash.edu.au/ jwb/edict_doc.html

4.1 Crosslingual Experiment

We compare our approach used for extracting cross-lingual translation pairs against several baselines. We compare to LLR + Manhattan (Rapp, 1999) and our variation LLR(P) + Manhattan. Additionally, we compare TFIDF(MSO) + Cosine, which is the TFIDF measure, whereas the Term Frequency is normalized using the maximal word frequency and the cosine similarity for comparison suggested in (Fung, 1998). Furthermore, we implemented two variations of this, TFIDF(MPO) + Cosine and TFIDF(MPO) + Jaccard coefficient, which were suggested in (Chiao and Zweigenbaum, 2002). In fact, TFIDF(MPO) is the TFIDF measure, whereas the Term Frequency is normalized using the maximal word pair frequency. The results are displayed in Figure 1. Our approach clearly outperforms all baselines; notably it has Top 1 accuracy of 0.14 and Top 20 accuracy of 0.55, which is much better than that for the best baseline, which is 0.11 and 0.44, respectively.

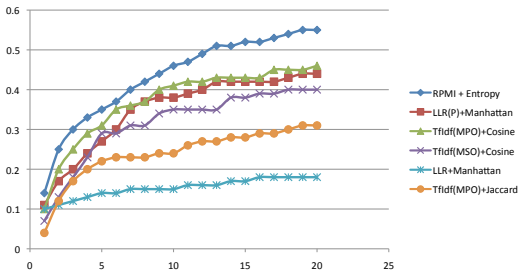

Figure 1: Crosslingual Experiment MLIT/NHTSA – Percentile Ranking of RPMI + Entropy Against Various Previous Suggested Methods.

We next leave the proposed framework constant, but change the mode of estimating positive associations and the way to match feature sets. As alternatives for estimating the probability that there is a positive association, we test LLR(P) and LLR. As alternatives for comparing feature sets, we investigate the matching coefficient (match), cosine similarity (cosine), Tanimoto coefficient (tani), and overlap coefficient

(over) (Manning and Schütze, 2002). The result of every combination is displayed concisely in Table 1 using the median rank[18]. The cases in which the median ranks are close to RPMI + Entropy are magnified in Table 2. We can see there that RPMI + Entropy, and LLR(P) + Entropy perform nearly equally. All other combinations perform worse, especially in Top 1 accuracy. Finally, LLR(P) presents a clear edge over LLR, which suggests that indeed only positive associations seem to matter in a cross-lingual setting.

	Entropy	Match	Cosine	Tani	Over
RPMI	13.0	17.0	24.0	37.5	36.0
LLR(P)	16.0	15.0	22.5	34.0	25.5
LLR	23.5	22.0	27.5	50.5	50.0

Table 1: Crosslingual experiment MLIT/NHTSA – Evaluation matrix showing the median ranks of several combinations of association and similarity measures.

	Top 1	Top 10	Top 20
RPMI + Entropy	0.14	0.46	0.55
RPMI + Matching	0.08	0.41	0.57
LLR(P) + Entropy	0.14	0.46	0.55
LLR(P) + Matching	0.08	0.44	0.55

Table 2: Accuracies for crosslingual experiment MLIT/NHTSA.

Finally we conduct an another experiment using the corpora pair Mainichi/Reuters which is quite different from MLIT/NHTSA. When comparing to the best baselines in Table 3 we see that our approach again performs best. Furthermore, the experiments displayed in Table 4 suggest that Robust PMI and pointwise entropy are better choices for positive association measurement and similarity measurement, respectively. We can see that

	Top 1	Top 10	Top 20
RPMI + Entropy	0.15	0.38	0.46
LLR(P) + Manhattan	0.10	0.26	0.33
TFIDF(MPO) + Cos	0.05	0.12	0.18

Table 3: Accuracies for crosslingual experiment Mainichi/Reuters – Comparison to best baselines.

[18] A median rank of i, means that 50% of the correct translations have a rank higher than i.

	Top 1	Top 10	Top 20
RPMI + Entropy	0.15	0.38	0.46
RPMI + Matching	0.08	0.30	0.35
LLR(P) + Entropy	0.13	0.36	0.47
LLR(P) + Matching	0.08	0.29	0.37

Table 4: Accuracies for crosslingual experiment Mainichi/Reuters – Comparison to alternatives.

the overall best baseline turns out to be LLR(P) + Manhattan. Comparing the rank from each word from the gold-standard pairwise, we see that our approach, RPMI + Entropy, is significantly better than this baseline in MLIT/NHTSA as well as in Mainichi/Reuters.[19]

4.2 Analysis

In this section, we provide two representative examples extracted from the previous experiments which sheds light into a weakness of the standard feature vector approach which was used as a baseline before. The two example queries and the corresponding responses of LLR(P) + Manhattan and our approach are listed in Table 5. Furthermore in Table 6 we list the pivot words with the highest degree of association (here LLR values) for the query and its correct translation. We can see that a query and its translation shares some pivots which are associated with statistical significance[20]. However it also illustrates that the actual LLR value is less insightful and can hardly be compared across these two corpora.

Let us analyze the two examples in more detail. In Table 6, we see that the first query ギア "gear"[21] is highly associated with 入れる "shift". However, on the English side we see that gear is most highly associated with the pivot word gear. Note that here the word gear is also a pivot word corresponding to the Japanese pivot word 歯車 "gear (wheel)".[22] Since in English the word gear (shift) and gear (wheel) is polysemous, the surface forms are the same leading to a high LLR value of

[19] Using pairwise test with p-value 0.05.

[20] Note that for example, an LLR value bigger than 11.0 means the chances that there is no association is smaller than 0.001 using that $LLR \sim \chi^2$.

[21] For a Japanese word, we write the English translation which is *appropriate in our context*, immediately after it.

[22] In other words, we have the entry (歯車, gear) in our dictionary but not the entry (ギア, gear). The first pair is used as a pivot, the latter word pair is what we try to find.

gear. Finally, the second example query ペダル "pedal" shows that words which, not necessarily always, but very often co-occur, can cause relatively high LLR values. The Japanese verb 踏む "to press" is associated with ペダル with a high LLR value – 4 times higher than 戻る "return" – which is not reflected on the English side. In summary, we can see that in both cases the degree of associations are rather different, and cannot be compared without preprocessing. However, it is also apparent that in both examples a simple L1 normalization of the degree of associations does *not* lead to more similarity, since the relative differences remain.

ギア "gear"		
Method	**Top 3 candidates**	**Rank**
baseline	jolt, lever, design	284
filtering	reverse, gear, lever	2
ペダル "pedal"		
Method	**Top 3 candidates**	**Rank**
baseline	mj, toyota, action	176
filtering	pedal, situation, occasion	1

Table 5: List of translation suggestions using LLR(P) + Manhattan (baseline) and our method (filtering). The third column shows the rank of the correct translation.

ギア		**gear**	
Pivots	**LLR(P)**	**Pivots**	**LLR(P)**
入る "shift"	154	gear	7064
入れる "shift"	144	shift	1270
抜ける "come out"	116	reverse	314
ペダル		**pedal**	
Pivots	**LLR(P)**	**Pivots**	**LLR(P)**
踏む "press"	628	floor	1150
戻る "return"	175	stop	573
足 "foot"	127	press	235

Table 6: Shows the three pivot words which have the highest degree of association with the query (left side) and the correct translation (right side).

5 Conclusions

We introduced a new method to compare context similarity across comparable corpora using a Bayesian estimate for PMI (Robust PMI) to extract positive associations and a similarity measurement based on the hypergeometric distribution (measuring pointwise entropy). Our experiments show that, for finding cross-lingual translations, the assumption that words with similar meaning share positive associations with the same words is more appropriate than the assumption that the degree of association is similar. Our approach increases Top 1 and Top 20 accuracy of up to 50% and 39% respectively, when compared to several previous methods. We also analyzed the two components of our method separately. In general, Robust PMI yields slightly better performance than the popular LLR, and, in contrast to LLR, allows to extract positive associations as well as to include prior information in a principled way. Pointwise entropy for comparing feature sets cross-lingually improved the translation accuracy clearly when compared with standard similarity measurements.

Acknowledgment

We thank Dr. Naoaki Okazaki and the anonymous reviewers for their helpful comments. Furthermore we thank Daisuke Takuma, IBM Research - Tokyo, for mentioning previous work on statistical corrections for PMI. This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan). The first author is supported by the MEXT Scholarship and by an IBM PhD Scholarship Award.

References

Chiao, Y.C. and P. Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the International Conference on Computational Linguistics*, pages 1–5. International Committee on Computational Linguistics.

Déjean, H., É. Gaussier, and F. Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the International Conference on Computational Linguistics*, pages 1–7. International Committee on Computational Linguistics.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Fung, P. 1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. *Lecture Notes in Computer Science*, 1529:1–17.

Garera, N., C. Callison-Burch, and D. Yarowsky. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 129–137. Association for Computational Linguistics.

Gaussier, E., J.M. Renders, I. Matveeva, C. Goutte, and H. Dejean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 526–533. Association for Computational Linguistics.

Haghighi, A., P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 771–779. Association for Computational Linguistics.

Johnson, M. 2001. Trading recall for precision with confidence-sets. Technical report, Brown University.

Koehn, P. and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, volume 34, pages 9–16. Association for Computational Linguistics.

Kudo, T., K. Yamamoto, and Y. Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230–237. Association for Computational Linguistics.

Manning, C.D. and H. Schütze. 2002. *Foundations of Statistical Natural Language Processing*. MIT Press.

Moore, R.C. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 333–340. Association for Computational Linguistics.

Moore, R.C. 2005. A discriminative framework for bilingual word alignment. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 81–88. Association for Computational Linguistics.

Morin, E., B. Daille, K. Takeuchi, and K. Kageura. 2007. Bilingual terminology mining-using brain, not brawn comparable corpora. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 45, pages 664–671. Association for Computational Linguistics.

Okazaki, N., Y. Tsuruoka, S. Ananiadou, and J. Tsujii. 2008. A discriminative candidate generator for string transformations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 447–456. Association for Computational Linguistics.

Pham-Gia, T. 2000. Distributions of the ratios of independent beta variables and applications. *Communications in Statistics. Theory and Methods*, 29(12):2693–2715.

Rapp, R. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Ross, T.D. 2003. Accurate confidence intervals for binomial proportion and Poisson rate estimation. *Computers in Biology and Medicine*, 33(6):509–531.

Tsuruoka, Y., Y. Tateishi, J. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *Lecture Notes in Computer Science*, 3746:382–392.

Wilcox, R.R. 2009. *Basic Statistics: Understanding Conventional Methods and Modern Insights*. Oxford University Press.