Bilingual lexicon extraction from comparable corpora: A comparative study

Nikola Ljubešić[1], Darja Fišer[2], Špela Vintar[2], Senja Pollak[2]
[1]Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000, Croatia
[2]Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000, Slovenia
E-mail: nljubesi@ffzg.hr, darja.fiser@ff.uni-lj.si, spela.vintar@ff.uni-lj.si, senja.pollak@ff.uni-lj.si

Abstract

This paper presents a comparative study of the impact of the key parameters for bilingual lexicon extraction for nouns from comparable corpora. The parameters we analyzed are: corpus size and comparability, dictionary size and type, feature selection for context vectors and window size, and association and similarity measures. Evaluation against the gold standard shows that window size of 7 with encoded position yields best results. The consistently best-performing association and similarity measures are Jensen-Shannon divergence with log-likelihood. We have shown that very good results can be achieved with small-sized but purpose-built seed lexicons and that problems arising from dissimilarities between the source and the target corpus can be compensated with their sufficient size.

1. Introduction

Bilingual lexica are the key component of all cross-lingual NLP applications and their compilation remains a major bottleneck in computational linguistics. Automatic extraction of translation equivalents from parallel texts has been shown extremely successful (e.g. Och and Ney, 2000; Tiedemann, 2005) but such a scenario is not feasible for all language pairs or domains because for many of them ready-made parallel corpora do not exist and their compilation is slow and expensive. This is why an alternative approach has been increasingly explored in the past decade that relies on texts in two languages which are not parallel but nevertheless share several parameters, such as topic, time of publication and communicative goal (Fung, 1998; Rapp, 1999). Compilation of such comparable corpora is much easier, especially since the availability of rich web data (Xiao & McEnery 2006).

In this paper we describe a set of experiments that serve to systematically determine the impact of the most important parameters for bilingual lexicon extraction from comparable corpora. The parameters we test and analyze are: the size and level of comparability of the corpus used for bilingual lexicon extraction; the type and size of the dictionary used to translate context vectors; the kind of features used to build context vectors and the amount of context that was taken into account; and, last but not least, the association and similarity measures used to compare the vectors across languages. The main contribution of this paper is a systematic comparison of various parameters that can serve as highly valuable guidelines on the collection of corpora and lexica for similar tasks.

The paper is structured as follows: in the next section we give an overview of previous work relevant for our research, Section 3 contains a description of the resources used and the steps taken in the experiment, in Section 4 we present the results of the evaluation of our approach and a discussion after which we conclude the paper with final remarks and ideas for future work.

2. Related work

For the task of bilingual lexicon extraction, parallel corpora provide very good results. However, the availability of parallel corpora is limited to certain language pairs and domains. Therefore, two main lines of research are proposed. The first one aims at bilingual lexicon extraction from comparable (non-parallel) corpora and the second one focuses on using the web to automatically construct parallel corpora (e.g. Fung et al., 2010). Our research falls in the first category.

The seminal papers in bilingual lexicon constructions are Fung (1998) and Rapp (1999) who proposed similar approaches that are based on the word co-occurrence hypothesis. Their main assumption is that the term and its translation share similar contexts. More recent adaptations of these approaches differ in the selection of methods at different stages.

Translation of vectors. At this stage, most researchers use machine-readable dictionaries. Some authors decide to prune out polysemous words in order to exclude semantic noise. Koehn and Knight (2002) build the initial seed dictionary automatically, based on identical spelling features. Cognate detection is used in a similar way by Saralegi et al. (2008), based on longest common subsequence ratio. Déjean et al. (2005), on the other hand, use a bilingual thesaurus instead of a bilingual lexicon.

Context representation. For selecting the representation of a word’s context, approaches differ mainly whether they look at a simple co-occurrence window of a certain size or decide to include some syntactic information as well. For example, Otero (2007) proposes binary dependences previously extracted from parallel corpus, while Yu and Tsujii (2009) use dependency parsers and Marsi and Krahmer use (2010) syntactic trees. Instead of context windows, Shao and Ng (2004) use language models.

Building feature vectors. The words in co-occurrence vectors can be represented as binary features, by term frequency or weighted by different association measures, such as TF-IDF (Fung, 1998), PMI (Shezaf and Rappoport, 2010) and, one of the most popular, the log-likelihood score. Others also investigate weighting co-occurrence terms differently if they appear closer to or further from the nucleus word in the context (e.g. Saralegi et al., 2008).

Selection of translation candidates. For ranking candidate translations, different vector similarity measures have been investigated. Rapp (1999) applies city-block metric, while cosine similarity (Fung, 1998) and Dice (Otero, 2007) seem to provide the best results. In addition, some approaches include re-ranking of translation candidates based on cognates detection (e.g. Saralegi, et al. 2008; Shao and Ng, 2004).

3. Experimental setup

In this section we give a detailed account of the experiments we conducted. In order to gain insight into the impact of the most important parameters for bilingual lexicon extraction, we ran a set of experiments in which we adjusted corpus size and the level of comparability of the texts between the languages. Next, we tested the translation of features in context vectors with three dictionaries of different type and size. Third, we tried out several settings of how to build context vectors and which association measure to use and finally, we tested different similarity measures to rank the translation candidates.

Although the parameters change in each run of the experiment, the basic algorithm for finding translation equivalents in comparable corpora is always the same:

(1) build context vectors for all unknown words in the source language and translate the vectors with a seed dictionary;

(2) build context vectors for all candidate translations words in the target language;

(3) compute the similarity for all translated source vectors and target vectors and rank translation candidates according to this score.

3.1 Corpora

Because it was our aim to analyze the impact of the size and comparability level of the corpus used to extract translation equivalents on the quality of the results we decided to use the English-Slovene part of the JRC-Acquis corpus (Steinberger et al., 2006). This is a 20-million-word parallel corpus of legislative texts, which we POS-tagged, lemmatized and filtered out punctuation and function words before we broke it into non-parallel corpora of different sizes and degrees of comparability.

We first took the English part of the corpus and sliced it into 10 equally-sized slices in chronological order, so that the first slice contained the oldest texts in the corpus and the last slice the most recent ones.

We then compared these slices with one another by computing the Spearman rank correlation coefficient (Kilgarriff, 2001) which compares the ranks of *n* most frequent words in each slice of the corpus. Such a comparison shows that slices from the same chronological period are more similar than those from different periods (e.g. the neighboring slices 3 and 4 are much more similar than the distant slices 2 and 9, see Table 1).

Now that we knew how similar or dissimilar these slices were, we were able to build several comparable corpora by taking the English part of the corpus for some slices and the Slovene part that corresponded to the other slices, making sure there was no overlaps between the slices used for one and the other language. In this way we built two sets of subcorpora; the first set consisted of subcorpora that contained slices with a high Spearman co-efficient, i.e. were highly comparable (called ‘easy1-5’ corpora), and the other set consisted of subcorpora populated with slices that had a low Spearman co-efficient, i.e. were not very comparable (called ‘hard1-5’ corpora). These two sets of subcorpora with very different levels of comparability were used to study the impact of corpora comparability on the quality of bilingual lexicon extraction.

Both sets of subcorpora consisted of 5 subcorpora, the smallest one containing a single slice per language (approx. 1.6 million content words) and the largest one 5 slices per language (approx. 8 million content words). The differently sized subcorpora were used to establish what is the smallest possible size of a comparable corpus that could still be used efficiently for finding translation equivalents.

High comparability (‘easy1-5’ corpora)			
Size	Slo slices	Eng slices	ρ
1.6	s3	s4	0.92
3.2	s1+s3	s2+s4	0.93
4.8	s1+s3+s5	s2+s4+s6	0.95
6.4	s1+s3+s5+s7	s2+s4+s6+s8	0.95
8	s1+s3+s5+s7+s9	s2+s4+s6+s8+s10	0.96
Low comparability (‘hard1-5’ corpora)			
Size	Slo slices	Eng slices	ρ
1.6	s2	s9	0.50
3.2	s1+s2	s9+s10	0.52
4.8	s1+s2+s3	s8+s9+s10	0.59
6.4	s1+s2+s3+s4	s7+s8+s9+s10	0.66
8	s1+s2+s3+s4+s5	s6+s7+s8+s9+s10	0.74

Table 1: Sets of subcorpora used in our experiment.

3.2 Dictionaries

In order to be able to compare vectors in different languages, a seed dictionary is needed to translate features in source context vectors. We tested our approach on three different dictionaries: a general large-sized bilingual dictionary (Grad), a medium-sized Wiktionary that covers basic vocabulary (Wiki), and a small domain-specific lexicon that was extracted from a word-aligned parallel corpus from the same domain (Acquis).

Only content-word dictionary entries were taken into account. No multi-word entries were considered either. And, since we do not yet deal with polysemy at this stage of our research, we only extracted the first sense for each dictionary entry. The seed dictionaries we obtained in this way contained from 2.800 entries (Acquis) to 6.600 entries (Wiki) and 42.700 entries (Grad).

A comparison of the extracted seed dictionaries with the JRC-Acquis corpus shows that even though the Grad dictionary is four times larger than the Acquis lexicon, the token overlap ratio is almost the same (81% vs. 78%). On the other hand, Wiktionary contains a similar amount of entries but they are not very relevant for the corpus in question (78% vs. 41%). We would like to see in our experiments whether reasonable results can be achieved with a small-sized lexicon with good coverage of the corpus vocabulary, so that large dictionaries which are difficult to obtain are no longer required.

	Types		Tokens	
Dict.	Overlap	Ratio	Overlap	Ratio
Grad	11,191	13.82%	5,634,190	81.73%
Wiki	3,122	3.86%	2,831,234	41.07%
Acquis	2,544	3.14%	5,401,254	78.35%

Table 2: A comparison of vocabulary coverage between the three dictionaries and the JRC-Acquis corpus.

3.3 Building and comparing context vectors

In this experiment we limited the task of extracting translation equivalents to nouns only, so we built context vectors for all those nouns that appear in the corpus at least 100 times and have at least 200 features (content words) in their context. We tested different window sizes (5, 7 and 9 lemmas). We compared two settings for feature selection: plain co-occurrence counts (i.e. bag-of-words approach) vs. included information on the position in which a context word appeared (e.g. L3-L2-L1-target_word-R1-R2-R3). With these settings, we extracted 1,105 vectors from the smallest subcorpus up to 2,494 vectors from the largest one.

In this way, we built vectors for all nouns in the source language and for all nouns in the target language. We tested four different association measures to represent features in the vector: relative frequency, pointwise mutual information (PMI), TF-IDF and log-likelihood (LL). Three variations of TF-IDF were taken into consideration: TF-IDF as defined in the information retrieval community (Spärck Jones, 1972), TF-IDF as defined in (Fung, 1998) and Okapi BM25 as the improved baseline in information retrieval (Robertson, 1994). Since none of the variations showed any significant difference, we disregarded the latter two.

Next, we translated words that appeared as features in the source context vector with a seed dictionary (see Section 3.2). If a feature word was not found in the dictionary, it was discarded from the context vector.

As the final step, the translated source context vector was compared to all target context vectors and the translation candidates were ranked according to their similarity score. The similarity measures we explored are: Manhattan and Euclidean distance, Jaccard and Dice indices adapted to non-binary values (Grefenstette, 1994), Tanimoto index (Tanimoto, 1957), cosine similarity and Jensen-Shannon divergence (Lin, 1991).

4. Evaluation

4.1 Automatic evaluation

Evaluation of the results was performed against a gold standard lexicon that was obtained from automatic word-alignment of a parallel corpus from the same domain. In the gold standard, there are several possible translations for the same source word, and we consider any of the variations as an equally suitable translation. The gold standard contains at least one translation for 1,000 source words.

Below we present the results of three experiments that best demonstrate the performance and impact of the key parameters for bilingual lexicon extraction from comparable corpora that we were testing in this research. The evaluation measure used throughout this research is mean reciprocal rank (Vorhees, 2001) on first ten candidates.

We start with the results for the largest subcorpus with a low comparability score (the hard5 subcorpus). The best-performing features for building context vectors turned out to be window size of 7 with encoded position of context words. The best-performing seed dictionary for translating vectors was the Acquis dictionary which was obtained from a small domain-specific word-aligned parallel corpus.

The measure that underperformed drastically on a regular basis under this setting was the Euclidean distance and was therefore removed from the rest of the experiments. Additionally, Dice gave consistently identical candidate lists as Jaccard and was therefore removed from the experiments as well.

The mean reciprocal rank scores for the described measures are given in Table 3. The best-performing combination is Jensen-Shannon divergence with log-likelihood, followed by Jaccard with log-likelihood and TF-IDF.

	relfreq	pmi	tfidf	ll
manh	0.07	0.11	0.15	0.04
jacc	0.70	0.62	0.74	0.74
tanim	0.57	0.49	0.60	0.43
cos	0.60	0.46	0.61	0.44
jenshan	0.68	0.51	0.69	0.78

Table 3: Evaluation of the results for different association and similarity measures on the hard5 subcorpus.

To get a better insight into the relationship between specific similarity measures and association measures, a series of visualizations is given. First, different similarity measures are compared on a boxplot in Figure 1. The variation in the data comes from using different association measures.

Manhattan is obviously overall the weakest similarity measure for this task while Tanimoto and cosine are regularly outperformed by Jaccard and Jensen-Shannon. Jaccard has more consistent results and could be considered the similarity measure of choice if one disregards the difference in association measures.

Additionally, different association measures are compared in the boxplot in Figure 2. Here the source of variation is different results obtained by different similarity measures. Pointwise mutual information obviously underperforms on a regular basis. Relative frequency, TF-IDF and log-likelihood obtain similar results. The variance in log-likelihood is much higher than in the other two association measures which shows its obvious sensitivity to different similarity measures.

In Figure 3 the same association measures are shown, but only for the two best performing similarity measures: Jaccard and Jensen-Shannon. Here the difference between the three best performing association measures becomes clearer. Log-likelihood is the best performing measure, whilst the second best is TF-IDF. The reason for relative frequency to perform that well in our opinion is the fact that the co-occurrence vectors are built from content words only and association measures do not play such an important role as would be if feature selection was less prohibitive.

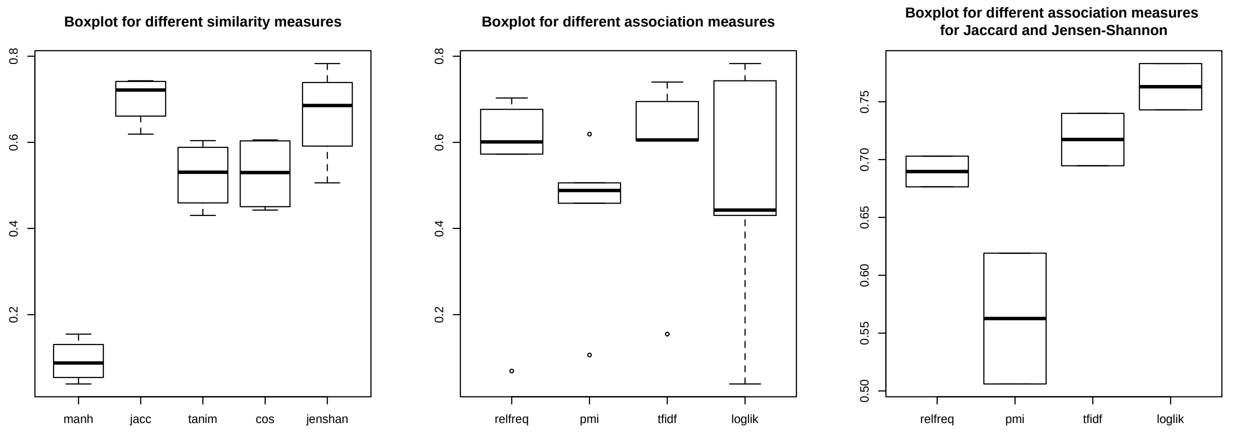
To analyze the consistency of the results, another two experiments were performed under different settings. This time, the smallest (easy1) and the largest (easy5) subcorpora with high comparability scores were used to obtain translation equivalents. These are, as stated before, built from more similar documents than the large, less comparable subcorpus (hard5). Additionally, the easy1 is five times smaller than the easy5 and hard5. In these two experiments, the Grad seed dictionary was used in the vector translation process as opposed to the prior experiment where the Acquis lexicon was used. The Pearson correlation coefficients between the results on hard5 on one side and easy1 and easy5 on the other side are computed. The results are given in Table 4.

	easy1	easy5
all values	0.975	0.982
association measures	0.912	0.957
similarity measures	0.997	0.999

Table 4. Correlation between the results on corpora easy1 easy5 with dict-grad and hard5 with dict-acquis.

The results show a high correlation between all results regardless of the resources and parameters used. When calculating the correlation of different association measure averages, the correlation decreases. On the contrary, when calculating the correlation between results on similarity measure averages, the correlation increases. These results show that specific similarity measures in general have more consistent results regardless of the experiment setting whereas association measures tend to show less consistency. We can conclude that the results of experiments with different settings are highly consistent with association measures being the cause for small variation.

The last experiment we wish to discuss here included different corpus sizes and degrees of comparability. As can be seen in Figure 4, the level of comparability of the corpora plays a major role in the quality of the extracted translation lexicon, especially when very little data is used. However, the size of the corpus is only significant with less comparable corpora. This is a very important finding because corpora with lower degrees of comparability are a much more likely scenario than nearly parallel ones, and it is encouraging to see that by simply increasing their size we can achieve results that are competitive with those obtained from nearly parallel corpora. It must be noted here that since we are using slices of a parallel corpus in this experiment, the level of comparability inevitably increases with corpus size, which is why a similar experiment should be conducted on real comparable corpora in order to confirm our findings in this research.



Figures 1-3: Visualization of the relationships between association and similarity measures regarding the mean reciprocal rank.

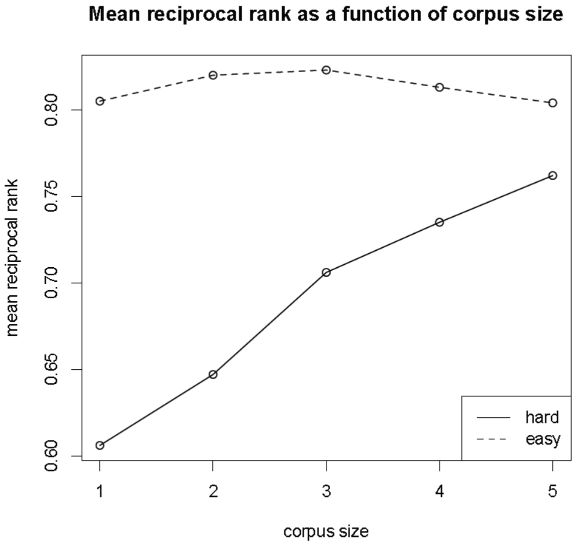


Figure 4. The impact on corpus size and comparability level.

4.2 Manual evaluation

For a more qualitative manual evaluation we selected 100 random source words from the hard5 corpus for which at least one translation candidate was generated, and examined the top ten translation equivalents for each word proposed by our system using the best-performing parameters. In 81 cases the first proposed equivalent matched at least one of the equivalents specified in the gold standard, whereby quite often the list of the extracted equivalents contained all the matches from the gold standard. In 4 cases where the first translation did not match the gold standard we saw that the proposed translation was in fact correct and that the gold standard could have been amended, for example (the correct equivalents are marked in bold):

source word: **integration**

gold standard: povezovanje, vključevanje

proposed equivalents (highest- to lowest-ranking):

integracija	**1.42** (missing in gold standard)
vključevanje	**1.56** (found in gold standard)
povezovanje	**1.59** (found in gold standard)
skupnost	1.64
dialog	1.65
razvoj	1.65
kohezija	1.66
partner	1.66
razsežnost	1.68
sodelovanje	1.69

In 14 cases the correct equivalent was not ranked first and these are the cases we plan to focus on in our future work; we believe that reranking methods applied at the post-processing stage could yet improve these results.

5. Conclusion

In this paper we described a set of experiments we conducted to gain more insight into what really matters in bilingual lexicon extraction for nouns from comparable corpora. The results show that window size of 7 with encoded position of context words are best settings for building context vectors. Small-sized domain specific lexicons that have good coverage of the vocabulary in the corpus can already achieve satisfactory results. This finding justifies the following research scenario as both feasible and efficient: first, a small parallel corpus in the relevant domain is compiled and word-aligned so that a seed lexicon is obtained, and then a much larger comparable corpus in the same domain is used for an extensive extraction of translation equivalents based on the seed lexicon.

What is more, we were able to show that a good combination of an association and similarity measure plays a much bigger role than feature selection or window size. The best-performing combination of association and similarity measures was consistently Jensen-Shannon divergence and log-likelihood. It is interesting to note that while log-likelihood is one of the most popular and best-performing similarity measures in the related work, Jensen-Shannon, which in our experiments outperforms the most popular cosine similarity measure and Dice coefficient, is on the other hand not used as an association measure in any related work we studied. A comparison of corpora of different sizes and degrees of comparability showed that for reasonable results, corpora do not necessarily need to be very similar since the lack of comparability can be compensated to a certain extent with a larger size.

In the future, we wish to test the approach on different corpora, domains and language pairs. In addition, we plan to look at various possibilities to rerank the translation candidates by taking into account cognates and named entities. We also wish to extend our work to other parts of speech and address polysemy as well as multi-word expressions.

6. Bibliography

Déjean, H., Gaussier, E., Renders, J.-M. and Sadat, F. (2005). Automatic processing of multilingual medical terminology: Applications to thesaurus enrichment and cross-language information retrieval. *Artificial Intelligence in Medicine*, 33(2): 111–124.

Fung, P. (1998). A statistical view on bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas*, pp. 1–17.

Fung, P., Prochasson, E. and Shi, S. (2010). Trillions of Comparable Documents. In *Proceedings of the 3rd workshop on Building and Using Comparable Corpora* (BUCC'10), Language Resource and Evaluation Conference (LREC2010), Malta, May 2010, pp. 26–34.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), pp. 97-133.

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proceedings of the workshop on Unsupervised lexical acquisition* (ULA'02) at ACL 2002, Philadelphia, USA, pp. 9–16.

Marsi, E. and Krahmer,E. (2010). Automatic analysis of semantic similarity in comparable text through syntactic tree matching. In *Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010), pp. 752–760.

Och, F. J. and Ney, H. (2000). Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics,* Hongkong, China, pp. 440–447.

Otero, P. G. (2007). Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In *Proceedings of the Machine Translation Summit* (MTS 2007), pp. 191–198.

Rapp, R. (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* (ACL '99), pp. 519–526.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. and Gatford, M .(1994) Okapi at TREC-3. *In Proceedings of the Third Text REtrieval Conference (TREC 1994).* Gaithersburg, USA.

Saralegi, X., San Vicente, I. and Gurrutxaga, A. (2008). Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *Proceedings of the 1st Workshop on Building and Using Comparable Corpora* (BUCC) at LREC 2008.

Shao, L. and Ng, H. T. (2004). Mining New Word Translations from Comparable Corpora. *In Proceedings of the 20th International Conference on Computational Linguistics* (COLING '04), Geneva, Switzerland.

Shezaf, D. and Rappoport, A. (2010). Bilingual Lexicon Generation Using Non-Aligned Signatures. In Proceedings *of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL 2010), Uppsala, Sweden, pp. 98–107.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1): 11–21.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), pp. 2142–2147.

Tanimoto, T.T. (1957) IBM Internal Report.

Tiedemann, J. (2005). Optimisation of Word Alignment Clues.*Natural Language Engineering*, 11(03), pp. 279–293.

Vorhees, E. M. (2001). *Overview of the TREC-9 Question Answering Track*. In Proceedings of the 9th Text Retrieval Conference (TREC-9) .

Xiao, Z. and McEnery, A. (2006). Collocation, semantic prosody and near synonymy: a cross-linguistic perspective. *Applied Linguistics* 27(1), pp. 103–129.

Yu, K. and Tsujii, J. (2009). *Bilingual dictionary extraction from Wikipedia.* In *Proceedings of the 12th Machine Translation Summit* (MTS 2009), Ottawa, Ontario, Canada.