Natural Language Engineering **11** (3): 327–341. © 2005 Cambridge University Press
doi:10.1017/S1351324905003876 Printed in the United Kingdom

Automatic bilingual lexicon acquisition using random indexing of parallel corpora

M. SAHLGREN and J. KARLGREN

Swedish Institute of Computer Science, SICS, Box 1263, SE-164 29 Kista, Sweden
e-mail: {mange, jussi}@sics.se

(*Received May 1 2004; revised November 30 2004*)

Abstract

This paper presents a very simple and effective approach to using parallel corpora for automatic bilingual lexicon acquisition. The approach, which uses the Random Indexing vector space methodology, is based on finding correlations between terms based on their distributional characteristics. The approach requires a minimum of preprocessing and linguistic knowledge, and is efficient, fast and scalable. In this paper, we explain how our approach differs from traditional cooccurrence-based word alignment algorithms, and we demonstrate how to extract bilingual lexica using the Random Indexing approach applied to aligned parallel data. The acquired lexica are evaluated by comparing them to manually compiled gold standards, and we report overlap of around 60%. We also discuss methodological problems with evaluating lexical resources of this kind.

1 Lexical resources should be dynamic

Lexical resources are necessary for any type of natural language processing and language engineering applications. Where in the early days of language engineering lexical information may have been hard-coded into the system, today most systems and applications rely on explicitly introduced and modularly designed lexica to function: examples range from applications such as automatic speech recognizers, dialogue systems, information retrieval, and writing aids to computational linguistic techniques such as part-of-speech tagging, automatic thesaurus construction, and word sense disambiguation systems.

Multilingual applications, which are driven by modeling lexical correspondences between different human languages, are obviously reliant on lexical resources to a high degree – the quality of the lexicon is the main bottleneck for quality of performance and coverage of service. Unfortunately, machine readable lexica in general, and machine readable multilingual lexica in particular, are difficult to come by. Manual approaches to lexicon construction vouch for high quality results, but are time- and labour-consuming to build, costly and complex to maintain, and inherently static: tuning an existing lexicon to a new domain is a complex task that risks compromising existing information and corrupting usefulness for previous application areas. Automatic lexicon acquisition techniques, on the other hand,

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

promise to provide fast, cheap and dynamic alternatives to manual approaches, but typically require sizeable computational resources and have yet to prove their potential in practical application.

This paper introduces a simple and effective approach to using distributional statistics over parallellized bilingual corpora for automatic multilingual lexicon acquisition. The approach is efficient, fast and scalable, and is easily adapted to new domains and to new languages. We evaluate the proposed methodology by first extracting bilingual lexica from aligned Swedish–Spanish and English–German data, and then comparing the acquired lexica to manually compiled gold standards. The results clearly demonstrate the viability of the approach.

2 Cooccurrence-based bilingual lexicon acquisition

Cooccurrence-based bilingual lexicon acquisition models typically assume something along the lines:

> "If we disregard the unassuming little grammatical words, we will, for the vast majority of sentences, find precisely one word representing any one given word in the parallel text. Counterterms do not necessarily constitute the same part of speech or even belong to the same word class; remarkably often, corresponding terms can be identified even where the syntactical structure is not isomorphic." (Karlgren 1988)

or alternatively formulated:

> "... words that are translations of each other are more likely to appear in corresponding bitext regions than other pairs of words." (Melamed 2000)

These models, first implemented by Brown, Cocke, Della Pietra, Della Pietra, Jelinek, Mercer and Roossin (1988), use aligned parallel corpora, and establish a translational relation between terms that are observed to occur with similar distributions in corresponding text segments. The calculation of correspondence between segments is an art in itself: translation is not typically done clause-by-clause, for reasons ranging from informational flow to stylistic finesse, to cultural differences, and to translator preference; arguably cannot be done term-by-term by virtue of the characteristics of human language; and seldom are consistent and correct for any longer stretch of texts due to limited stack space of human translators.

Given a set of text segments in two (or more) languages – here arbitrarily called the source and the target languages[1] – that have been aligned satisfactorily, correspondences between term occurrences across the source and target texts can be calculated in several ways. Most approaches calculate *cooccurrence scores* between terms using various related formulas – the number of times a term w_s from the source language occurs in a source text segment $T_s(i)$ folded together with the number of times a candidate term w_t occurs in the corresponding text segment in the target language $T_t(i)$. These scores are taken as the primary datum, and similarities between terms are then calculated using these scores as a base.

[1] Naturally, nothing precludes these two sets of text segments from being written in the *same* language.

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

There are major drawbacks to this approach. Most importantly, as has been noted by several practitioners in the field, the assumption that translation involves correspondences between single lexical items is a patent oversimplification. While the operationalization of translation as a correspondence between a lexical item in the source language with an item in the target language certainly is convenient for present purposes, the assumption that meaning is compositional from morpheme level upwards is at the least questionable. For us, the primary meaning bearing unit is the utterance, the coherent expression of something meaningful by a speaker or a writer. Our model does not make explicit use of cooccurrence between lexical items in corresponding languages, but rather of individual occurrences of lexical items in contexts of use. They model utterances, not lexical items – in the bilingual case, translated pairs of utterances.

Confounding the model by trying to model collocational data separately from cooccurrence across languages does not improve its theoretical basis. Modeling occurrence data is basic; lexical items that occur in the same clause *within* one language are indelibly related through that syntagmatic relation – however the relation is modeled by linguists – and the entire utterance bears a relation to the translation of it in the target language. In this sense, *every* term of the source utterance is related to *every* term in target utterance, even if their relative import may differ by orders of magnitude. Unrefined term occurrence data are, however, unsatisfactory means to model semantic relations because terms are polysemous and have synonyms as a matter of course, and a matrix of term-by-term relations will mostly contain empty cells, modeling nothing but lack of relevant observations. Thus the high dimensionality of such models will, through its low level of abstraction, obscure semantically salient dependencies – the space of semantic variation appears in some sense to be of a lower inherent dimensionality.

Aside from underlying assumptions of compositionality and distributional semantics, methodologies that take term cooccurrences as primary data will have to address practical issues that have little to do with meaningful semantic correspondence. They will inevitably run into scalability and tractability problems when presented with real data comprising hundreds of thousands of term tokens in millions of texts. These problems will compound all the more rapidly when the data are multilingual (Gale and Church 1991).

3 Context-based bilingual lexicon acquisition

Our approach, by contrast, takes the *context* – an utterance, a window of adjacency, or when necessary, an entire document – as the primary unit. Rather than building a huge vector space of contexts by lexical item types, as most retrieval systems do, implicitly or explicitly, we build a vector space which is large enough to accommodate the occurrence information of tens of thousands of lexical item types in millions of contexts, yet compact enough to be tractable; constant in size in face of ever-growing data sizes; and designed to model association between distributionally similar lexical items without compilation or explicit dimensionality reduction.

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

Table 1. *Example of parallel data*

T_s	Context	T_t
a a a b c c	1	x v y z z z z
a d e	2	v w z
a a a c	3	x x v

To illustrate the difference between purely cooccurrence-based approaches and our context-based approach, consider Table 1.

Brown's original model (Brown *et al.* 1988; Brown, Cocke, Della Pietra, Della Pietra, Jelinek, Lafferty, Mercer and Roossin 1990) and Melamed's later model (Melamed 2000), to take two examples, differ in how they calculate the cooccurrence of w_s and w_t. Brown's cooccurrence measure is proportional to the product of their joint frequencies in each step; Melamed's dampens the measure by only taking the smaller of the frequencies. In this case, Brown's model would find *a* and *z* having the closest cooccurrence score – with the score for context 1 dominating everything else; Melamed's model would find *a* and *v* to be the closest, with the number of contexts they engage in dominating everything else; our model will – under typical parameter settings – find that *a* has the closest cooccurrence score with contexts 1 and 3, that *x* also has the same context profile, and that they thus are the most closely corresponding terms.

In other similar experiments, other term distribution measures have been used to temper and modulate the effects of the pure cooccurrence measure. Similarity metrics that weight together raw cooccurrence with global occurrence statistics (under the assumption that a term that occurs often elsewhere in other contexts is a bad candidate); term length in characters, (under the assumption that semantically similar terms tend to have similar graphotactic appearance); term position in the respective text segments (under the assumption that source and target languages have similar syntactic characteristics) have all been tested, often usefully (Karlgren, Karlgren, Nordström, Pettersson and Wahrolén 1994) – but these different types of information sources often require weeding out weak translation candidates using filters such as other lexical resources, e.g. based on lexical categorization of terms based on their part-of-speech. Currently, we do not implement such filters, in keeping with our principle of association being a relation between terms and utterances rather than between terms and terms.

4 Random indexing

The context-based approach that we propose is based on *Random Indexing* (Kanerva, Kristofferson and Holst 2000; Karlgren and Sahlgren 2001). Random Indexing is a methodology for producing *context vectors* that represent the distributional profile of linguistic entities. The rationale of context vectors is that they make it possible to compute distributional similarity between linguistic entities by using standard vector

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

similarity measures (Gallant 1991). Context vectors are normally produced by using the standard vector space model, which represents the text data in a cooccurrence matrix F of order $w \times d$, such that the rows F_w represent the terms, the columns F_d represent the documents (or whatever text section one wants to use as context), and the cells are the (possibly weighted and normalized) frequency of a given term in a given document. Each row of frequency counts thus constitutes a d-dimensional context vector $\vec{w}$ for a given term.

A serious problem with the standard vector space model is the dimensionality d of the context vectors, which often is very large, and which will continue to increase whenever new data are added. Very high (and increasing) dimensionality is a severely limiting factor with regards to scalability, efficiency, and applicability of the vector space methodology. In many real world applications, it is therefore necessary to use dimension reduction techniques.[2] Unfortunately, dimension reduction can be a computationally expensive operation, and it is typically applied only after the initial huge $w \times d$ matrix has been assembled. Thus, if new data are encountered, the entire process of first building matrix F and then reducing it has to be repeated from scratch.[3]

Random Indexing overcomes these scalability and efficiency problems by *incrementally* accumulating k-dimensional index vectors into a context matrix G of order $w \times k$, where $k \ll d$. This is done in a two-step operation:

1. Each context (e.g. each document or each position in a sliding window of word tokens) in the text is assigned a unique and randomly generated representation called an *index vector*. These index vectors are sparse, high-dimensional, and ternary, which means that their dimensionality k is on the order of thousands, and that they consist of a small number (ϵ) of randomly distributed +1s and −1s, with the rest of the elements set to 0.
2. Context vectors are produced by scanning through the text, and each time a term occurs in a context (e.g. document), that context's k-dimensional index vector is added to the row for the term in matrix G. Terms are thus represented in the context matrix by k-dimensional context vectors that are effectively the sum of the terms' contexts.

Note that the same procedure will produce a standard cooccurrence matrix F of order $w \times d$ if we use unary index vectors of the same dimensionality as the number of contexts.[4] Such d-dimensional unary vectors would be orthogonal, whereas the k-dimensional random index vectors are only *nearly* orthogonal. However, as Hecht-Nielsen has shown (Hecht-Nielsen 1994), there are many more nearly orthogonal

[2] Commonly used dimension reduction techniques include factor analytic methods such as singular value decomposition (used in Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer and Harshman 1990; Landauer and Dumais 1997)), and more linguistically motivated term filtering methods.

[3] Strictly speaking, even LSA can be incrementally extended, although it is computationally very expensive.

[4] These unary index vectors would have a single 1 in a different position for each context, i.e. the nth context will have an index vector of length d, with a single non-zero element in the nth position.

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

than truly orthogonal directions in a high-dimensional space. Thus, choosing random directions gets us sufficiently close to orthogonality, which means that the k-dimensional random index vectors *approximate* the d-dimensional unary vectors. Consequently, context matrix $G_{w\times k}$ will be an approximation of cooccurrence matrix $F_{w\times d}$ in the sense that their corresponding rows are similar or dissimilar to the same degree.

The near-orthogonality of random directions in a high-dimensional space is the key to a family of dimension reduction techniques that includes methods such as Random Projection (Papadimitriou, Raghavan, Tamaki and Vempala 1998), Random Mapping (Kaski 1999), and Random Indexing. These methods rest on the same insight – the Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss 1984), which states that if we project points in a vector space into a randomly selected subspace of sufficiently high dimensionality, the distances between the points are approximately preserved. Thus, the dimensionality of a given matrix F can be reduced by multiplying it with (or projecting it through) a random matrix R:

$$G_{w\times k} = F_{w\times d}R_{d\times k}$$

Obviously, the choice of random matrix R is an important design decision for dimension reduction techniques that rely on the Johnson–Lindenstrauss lemma. As we saw above, if the d random vectors in matrix R are orthogonal, so that $R^TR = I$, then $G = F$; if the random vectors are nearly orthogonal, then $G \approx F$ in terms of the similarity of their rows. A very common choice for matrix R is to use Gaussian distribution for the elements of the random vectors. However, Achlioptas has shown that much simpler distributions – practically all zero mean distributions with unit variance – give a mapping that satisfies the lemma (Achlioptas 2001). Random Indexing is equivalent to Achlioptas' proposal with parameters:

$$r_{ij} = \frac{1}{\epsilon}\sqrt{k} \times \begin{cases} +1 & \text{with probability } \frac{(\epsilon/2)}{k} \\ 0 & \text{with probability } \frac{k-\epsilon}{k} \\ -1 & \text{with probability } \frac{(\epsilon/2)}{k} \end{cases}$$

where k is the dimensionality of the vectors, and ϵ is the number of non-zero elements in the random index vectors.

4.1 Advantages of random indexing over other similar models

Compared to other vector space methodologies, the Random Indexing approach is unique in the following three ways.

First, it is an *incremental* method, which means that the context vectors can be used for similarity computations even after just a few examples have been encountered. By contrast, most other vector space methods require the entire data to be sampled and represented in a huge cooccurrence matrix before similarity computations can be performed.

Second, it uses fixed dimensionality, which means that new data do not increase the dimensionality of the vectors. Increasing dimensionality can lead to significant scalability problems in other vector space methods.

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

Third, it uses implicit dimension reduction, since the fixed dimensionality is much lower than the number of contexts in the data. This leads to a significant gain in processing time and memory consumption as compared to vector space methods that employ computationally expensive dimension reduction algorithms.

As an example, the complexity of computing a singular value decomposition (as is done in LSA) is on the order of $O(wzd)$ (under the assumption that the data are sparse), where w is the size of the vocabulary, d is the number of documents, and z is the number of non-zero elements per column (Papadimitriou *et al.* 1998). Performing a random projection of the original (sparse) data (i.e. forming a $w \times k$ random matrix and projecting the original $w \times d$ matrix through it) is $O(zkw)$ (Papadimitriou *et al.* 1998; Bingham and Mannila 2001), where k is the dimensionality of the vectors. By contrast, producing context vectors with Random Indexing is only $O(wk)$, since it is not reliant on the initial construction of the huge cooccurrence matrix.

4.2 Random indexing for context-based bilingual lexicon acquisition

Applying the Random Indexing approach to the present cross-lingual application is done by first assigning one random index vector to each aligned pair of text segments in the source and target languages. We then produce context vectors for the terms in both languages by adding an aligned segment's index vector $\vec{a}$ to the context vector $\vec{w}$ for a given term every time the term occurs in the aligned segment:

$$\vec{w} = \sum_{w \in a} \vec{a}$$

This means that the terms from the source language and the target language are both located in the same vector space, effectively constituting a bilingual lexicon, where translations are defined as those terms in the different languages whose context vectors are most similar. Similarity can be straightforwardly computed using standard distance metrics from vector space algebra. In these experiments, we use the cosine of the angles between the context vectors, defined as:

$$d_{cos}(x, y) = \frac{\vec{x} \cdot \vec{y}}{| \vec{x} || \vec{y} |} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

5 Evaluation methodology

Evaluation of a lexicon is of course best done through comparison with an existing lexicon of the same type. For digital lexical resources this involves a bootstrapping problem: if no such resource exists something needs to be used provisionally. In our case, we resort to bilingual lexica intended for human use, with limited yet sizeable coverage: we take for each source-language term w_s the target-language term w_t our system finds as the closest neighbor and compare it to the various terms these lexica give as translations for w_s. In our case there is an important mismatch between the gold standard lexica and the type of results our system produces – our system is not intended to give one-to-one correspondences on a term level.

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

To simplify comparisons, we use roughly the same experimental setup as Sahlgren in previous experiments (Sahlgren 2004). The main difference here is that our evaluation is based on the entire available data set rather than on a random sample.

Freely available parallel texts are used as data, and freely available bilingual lexica are used as gold standards. The evaluation procedure simply consists in computing the overlap between the automatically acquired lexicon and the gold standard. In the following sections, we describe the data and the evaluation metric.

5.1 Data

We use the document-aligned Swedish–Spanish and English–German Europarl corpora (Koehn 2002),[5] which we lemmatize using tools from Connexor.[6] The resulting vocabularies are 100,891 terms in Swedish, 42,761 terms in Spanish, 40,181 terms in English, and 70,384 terms in German. The Swedish–Spanish data consist of 37,379 aligned document pairs, and the English–German data consist of 45,556.

5.2 Deriving the bilingual lexicon

To extract bilingual lexica using the Random Indexing approach, we first define *source vocabularies* to consist of all one-word terms in the Swedish and English data that also occur in the gold standard lexica. We also define *target vocabularies* to include *all* Spanish and German terms, i.e. we do not restrict the target vocabularies according to the gold standard lexica; therefore, we may encounter target terms that are valid translations, but that will not be counted as such because they are not featured in the gold standard lexica.

Next, for each *source term* in the source vocabularies, we extract the highest correlated one-word term in the corresponding target vocabulary, and define this term as the *target term* (or *translation candidate*). Correlation between terms is computed as the cosine of the angles between the context vectors of the source and target terms. Note that the cosine metric gives the correlation between any two terms, and that this measure can be used as a confidence score for each translation candidate.

5.3 Evaluation metric

The quality of the derived lexica are measured by comparison against online lexical resources. For Swedish–Spanish, we use Lexin's online Swedish–Spanish lexicon,[7] while for English–German, we use TU Chemnitz' online English–German dictionary.[8] We look up each source term in the respective lexica and inspect the

[5] The Europarl corpora consist of parallel texts from the proceedings of the European Parliament, and is available in 11 European languages. The data are freely available at http://people.csail.mit.edu/people/koehn/publications/europarl/

[6] http://www.connexor.com

[7] http://lexin.nada.kth.se/sve-spa.shtml

[8] http://dict.tu-chemnitz.de/

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

lexicon entry to see if the target term occurs among the terms in the entry. We do not attempt analysis of compounds and multi-word terms but take each occurrence of a target term as a hit, regardless of which type of construction it participates in in the entry. We use *Precision* as our evaluation metric, defined as: $Precision = \frac{C}{S}$, where C is the number of correct entries in the acquired lexicon, and S is the size of the source vocabulary.

6 Experiments and results

Three sets of experiments are conducted. In the first set, we look at the relationship between a term's frequency and the quality of its translation. In the second set, we investigate the effects of using different dimensionalities of the vectors. In the third set, we examine how the results are influenced by increasing the number of translation candidates.

6.1 The effects of frequency

Grefenstette (1993) notes that automatic lexicon acquisition techniques are liable to frequency effects – that is, the methods tend to work better for terms with high and medium frequency. This is not very surprising, as high-frequency terms provide better statistics, and will therefore produce more reliable estimates than low-frequency terms (Grefenstette 1993). Sahlgren's previous experiments support this claim (Sahlgren 2004).

To investigate whether such a correlation between performance and frequency can be identified for the present data, we measured precision over 8 different source-term frequency ranges: 5–10, 10–50, 50–100, 100–500, 500–1,000, 1,000–5,000, 5,000–10,000, 10,000–500,000 with settings $k = 600$, $\epsilon = 6$.[9] The results are displayed in Figure 1.

As can be seen, there is a strong correlation between a term's frequency and the quality of its translation. Terms with high and medium frequency produce much better results than terms with low frequency; terms with a frequency below 100 do not produce reliable statistics and should therefore be excluded from the acquired lexica. These observations concord with those of Grefenstette (1993) and Sahlgren (2004).

Accordingly, we redefine our source vocabularies to only include terms that occur in the gold standard lexica, and that occur more than 100 times in the data. This reduces the source vocabularies from 100,904 to merely 4,855 unique terms in Swedish, and from 40,181 to 4,902 in English.

[9] We use a comparatively low dimensionality in these experiments for efficiency reasons. Other parameter settings might be optimal (see the next set of experiments), but since we are only interested in the relationship between frequency and performance here, optimality is not crucial – the correlation between frequency and performance will be the same regardless of the dimensionality of the vectors.

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

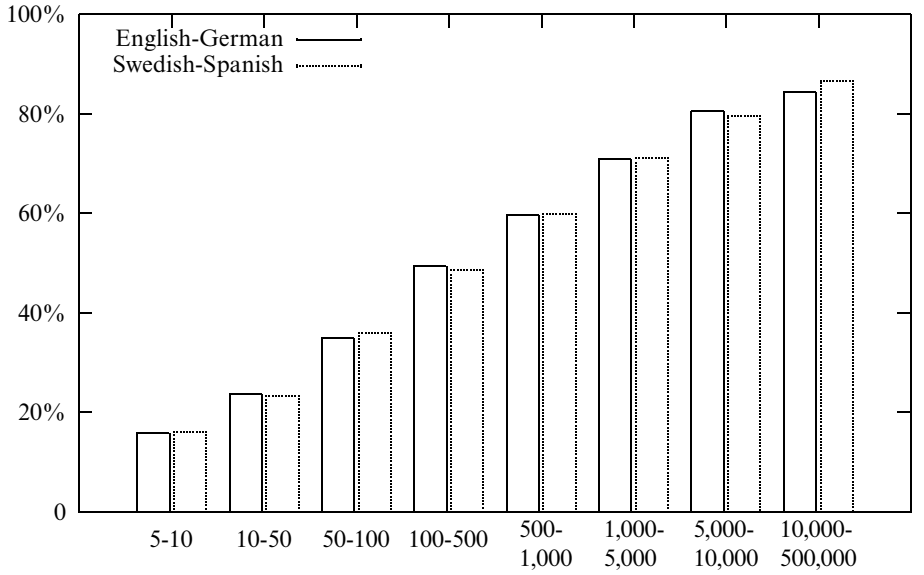


Fig. 1. Precision for eight different frequency ranges, with $k = 600$ and $\epsilon = 6$.

6.2 *The effects of dimensionality*

In theory, the random projection family of dimension reduction techniques should give a better approximation to the original data matrix as the dimensionality of the random matrix increases (e.g. Kaski (1999) shows that the higher the dimensionality of the random vectors, the closer the matrix $R^T R$ will approximate the identity matrix, and Bingham and Mannila (2001) observe that the mean squared difference between $R^T R$ and the identity matrix is about $\frac{1}{k}$ per element). In order to evaluate the effects of increasing the dimensionality in this application, we computed precision using 12 different dimensionalities of the vectors, with k ranging from 100 to 6,000 and ϵ ranging from 2 to 60 (depending on k). Note that a standard vector space model would require 37,379 dimensions for the Swedish–Spanish data, and 45,556 dimensions for the English–German data.

In these experiments, the results are reported using average precision over five different runs. This is done in order to counter the risk of noise from randomness – each new configuration of index vectors will produce slightly different results, since they are chosen at random. As suggested by the previous experiment, we only include words with frequency > 100 in the source vocabularies. The results are displayed in Figure 2.

This experiment demonstrates that the quality of the context vectors does increase with their dimensionality as expected, but that the gain levels off when the dimensionality becomes sufficiently large. A similar trend is observed by Sahlgren and Cöster in another application of Random Indexing (Sahlgren and Cöster 2004). There is hardly any difference in performance when the dimensionality of the vectors exceeds 3,000. The precision is approximately 60% even if the dimensionality

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

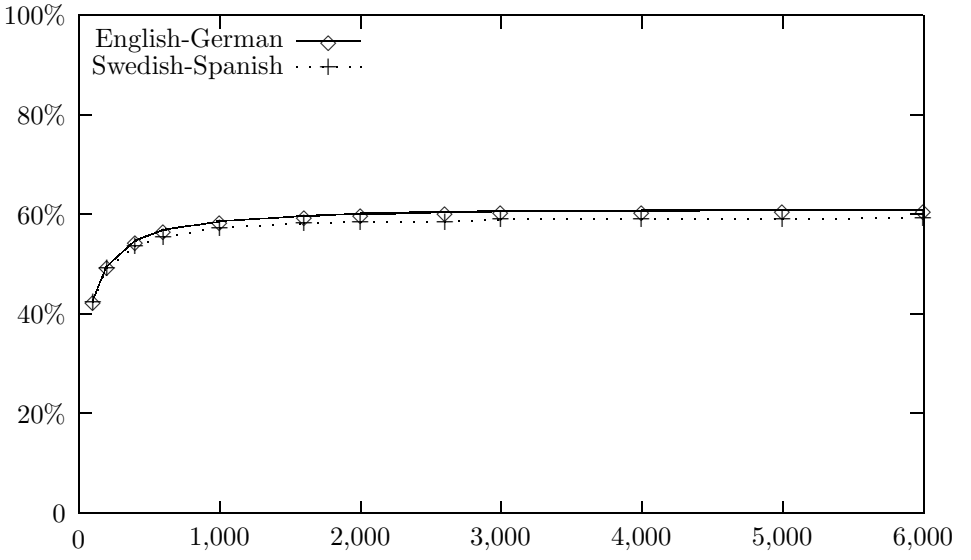


Fig. 2. Average precision over five runs for dimensionalities ranging from 100 to 6,000.

is increased from 3,000 to 6,000. However, there is a drastic decrease in performance when the dimensionality of the vectors drops below 1,000. This clearly demonstrates the importance of choosing a sufficiently high dimensionality for the vectors. At the same time, it is advisable to choose the lowest reliable dimensionality for efficiency reasons. A choice of 3,000 dimensions seems appropriate for the present data. Note that the original dimensionality of the data is more than 10 times as large.

6.3 The effects of lexicon size

The approach to lexicon acquisition presented in this paper is particularly well suited to handle situations were there might be more than one appropriate translation of a particular term. Since the multilingual vector space effectively relates all the terms in the different languages by the similarity of their context vectors, we can easily extend the lexicon with not only the most similar target term, but the m most similar terms.

In this set of experiments, we examine how the precision of the lexica is influenced when we extend the number of translation candidates in the lexicon entries to m, where $m = 1$, 2, 3, 4, 5, 10, and 20. Note that we use the same evaluation metric as in the other experiments, which means that it will be sufficient for *one* of the translation candidates in a lexicon entry to occur in the gold standard in order for the entry to count as correct, i.e. we do not require *all* the translation candidates to be correct in order to count the *entry* as correct. For efficiency reasons, this experiment uses $k = 600$ and $\epsilon = 6$, and only includes source terms with frequency exceeding 100.

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

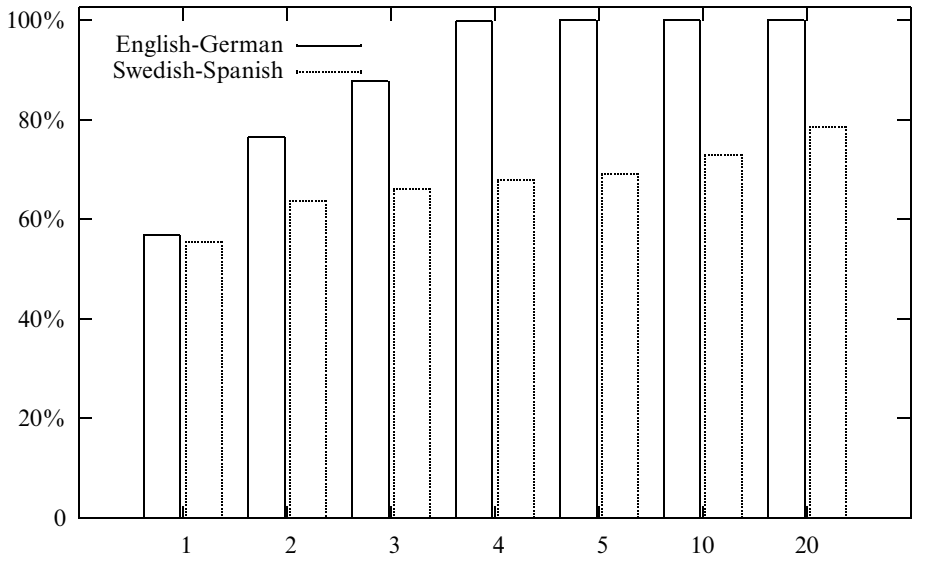


Fig. 3. Precision for different numbers of translation candidates in the English–German and Swedish–Spanish lexica.

Figure 3 clearly shows that while the first translation candidate is not always correct, appropriate translations can often be found a little further down the list. For the English–German data, 100% of the entries contain correct translations if more than three candidates are considered. This indicates that it might be possible to improve the quality of the acquired lexicon by reranking some of the translation candidates. It also suggests that in a situation where the automatically acquired lexica are used interactively by human users, e.g. as raw material for manual resource development, it would be advisable to include several translation candidates in the entries.

7 Towards a reliable and valid evaluation methodology

The experimental results vary as a function of both the quality of the parallel data used as input, and the coverage of the gold standard. If the parallel data are of low quality the results will be unsatisfactory irrespective of qualities of the algorithm; if the coverage of the gold standard is insufficient, the experimental results will not be reliable. The first of these factors points to the importance of creating, maintaining, and providing publicly available high-quality parallel data for the systematic evaluation of language engineering techniques and applications. The second factor points to the need for devising more pertinent and robust evaluation procedures for multilingual language engineering research.

As an example, to provide a comparison with other methods, we ran the same data and the same evaluation scheme using the GIZA++ statistical translation modeling package (Och and Ney 2000). As part of its output, GIZA++ suggests

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

best translations for words in the source text. GIZA++'s suggested best translations are correct up to something less than $\frac{1}{3}$. The comparison is not unproblematic to evaluate: while we are happy to see that our approach provides much better figures, GIZA++ does expect the data to be sentence aligned. Our method can be used with any kind or range of aligned text segments. On the one hand, this means that the comparison is less than adequate, and on the other hand, this shows how our approach can do better with less refined data. With a common evaluation framework the comparison could be less equivocal.

While the coverage of the gold standard used in these experiments is fair,[10] it is by no means complete with respect to the data used for these experiments. This means that a considerable amount of potentially correct and partially correct translations will not be found in the gold standards, and will therefore not be counted as correct translations in the evaluation. Examples of (potentially) correct translations from our automatically acquired lexica that were not featured in the gold standards include: "constantly"/"ständig" (from the English–German data), and "juridiskt"/"jurídicamente" (from the Swedish–Spanish data). Examples of partially correct translations are: "working"/"Arbeitsgruppe", "working"/"Arbeitszeit" (from the English–German data), and "socialförsäkring"/"social" (from the Swedish–Spanish data). The last examples demonstrate that the difference between compounding languages (such as Swedish and German) and non-compounding languages (such as Spanish and English) needs to be specifically addressed in this type of application.

Another type of error that is common in this kind of evaluation scheme is morphological and orthographical variation, where the manually compiled lexicon features a different spelling or a different morphological realization of a particular term. One example from our experiments is the entry "Jugoslavisk"/"Yugoslava" (from the Swedish–Spanish data), which is not counted as a correct translation because the Spanish gold standard only contains "Yugoslavo".

In order to get a fair assessment of the quality of the translation candidates that were not featured in the gold standard, we asked professional translators at a language consultancy company to manually review a random sample of 600 out-of-vocabulary translation candidates, and to assess whether they were correct. To investigate whether a more relaxed correctness criterion might be motivated, we asked the judges to grade each term pair as either a "perfect hit", "not irrelevant" or "miss". Approximately $\frac{1}{3}$ of the initially rejected translation candidates turned out to be appropriate translations (i.e. "perfect hits"), and should have been counted as correct in the evaluation. About another $\frac{1}{10}$ of the rejected translation candidates were graded as "not irrelevant".

8 Concluding remarks

The methodology presented in this paper provides a simple and effective approach to using parallel texts for automatic bilingual lexicon acquisition. The approach,

[10] Our versions of Lexin's Swedish–Spanish lexicon and TU Chemnitz' English–German dictionary contains 13,653 and 116,273 entries, respectively.

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

which uses the Random Indexing vector space methodology, requires a minimum of preprocessing and linguistic knowledge, and is efficient, fast and scalable. The approach is applicable to any domain and to any language, and does not require any external resources.

In this paper, we have shown how to apply the Random Indexing procedure to aligned parallel data, and how to extract bilingual lexica from the resulting vector spaces. The quality of the acquired lexica was evaluated by comparing them to manually compiled gold standards. The overlap was around 60%, when only terms with frequency above 100 occurrences in the source languages were included.

We conclude that, while there is a number of inherent problems with our evaluation procedure – counting the overlap between the automatically acquired lexica and manually compiled gold standards – the results demonstrate the viability of the proposed approach for automatic bilingual lexicon acquisition. The results are promising even in the face of a task not ideally suited for our theoretical model. Our belief is that our system is better suited for an interactive service for human professionals or experienced language users than for batch mode processing; evaluating such as service would be better done in terms of user acceptance than by precision measures of a fully automated task. In either case, our experiments certainly prove the viability of vector space models for multi-lingual tasks, including that of generating multi-lingual terminological resources.

Acknowledgements

The work reported here has been funded by the European Commission under contract IST-2000-29452 (DUMAS – Dynamic Universal Mobility for Adaptive Speech Interfaces). We wish to thank Pentti Kanerva and three anonymous reviewers for valuable comments. Thanks also to Connexor for providing morphological analysis of the data, and to Linguaweb, specialists in Spanish–Swedish language services, for assessing correctness of translation candidates.

References

Achlioptas, D. (2001) Database-friendly random projections. *Symposium on Principles of Database Systems*.

Bingham, E. and Heikki M. (2001) Random projection in dimensionality reduction: applications to image and text data. *Knowledge Discovery and Data Mining*.

Brown, P., Cocke, S., Della Pietra, V., Della Pietra, F., Jelinek, F., Mercer, R. and Roossin, P. (1988) A statistical approach to language translation. *Proceedings of the 12th Annual Conference on Computational Linguistics (COLING 88)*. International Committee on Computational Linguistics.

Brown, P. F., Cocke, J., Della Pietra, S., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roossin, P. S. (1990) A statistical approach to machine translation. *Computational Linguistics* **16**(2): 79–85.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R. (1990) Indexing by latent semantic analysis. *Journal of the Society for Information Science* **41**(6): 391–407.

Gale, W. and Church, K. (1991) Identifying word correspondences in parallel texts. *Proceedings of the DARPA Workshop on Speech and Natural Language*.

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press

Gallant, S. I. (1991) Context vector representations for document retrieval. *AAAI Natural Language Text Retrieval Workshop*.

Grefenstette, G. (1993) Evaluation techniques for automatic semantic extraction: Comparing syntactic and window-based approaches. *Workshop on Acquisition of Lexical Knowledge from Text*.

Hecht-Nielsen, R. (1994) Context vectors: general purpose approximate meaning representations self-organized from raw data. In: J. M. Zurada, R. J. Marks II and C. J. Robinson, editors, *Computational Intelligence: Imitating Life*, pp. 43–56. IEEE Press.

Johnson, W. B. and Lindenstrauss, J. (1984) Extensions of lipshitz mapping into hilbert space. *Contemporary Mathematics* **26**: 189–206.

Kanerva, P., Kristofersson, J. and Holst, A. (2000) Random indexing of text samples for latent semantic analysis. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Erlbaum.

Karlgren, H., Karlgren, J., Nordström, M., Pettersson, P. and Wahrolén, B. (1994) Dilemma – an instant lexicographer. *Proceedings of the 15th Annual Conference on Computational Linguistics (COLING 94)*. International Committee on Computational Linguistics.

Karlgren, H. (1988) Term-tuning, a method for the computer-aided revision of multi-lingual texts. *International Forum for Information and Documentation* **13**(2): 7–13.

Karlgren, J. and Sahlgren, M. (2001) From words to understanding. In: Y. Uesaka, P. Kanerva and H. Asoh, editors, *Foundations of Real-World Intelligence*, pp. 294–308. CSLI Publications.

Kaski, S. (1999) Dimensionality reduction by random mapping: Fast similarity computation for clustering. *Proceedings of the IJCNN'98, International Joint Conference on Neural Networks*. IEEE Service Center.

Koehn, P. (2002) Europarl: A multilingual corpus for evaluation of machine translation.

Landauer, T. and Dumais, S. (1997) A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* **104**(2): 211–240.

Melamed, D. (2000) Models of translational equivalence among words. *Computational Linguistics* **26**(2): 221–249.

Och, F. J. and Ney, H. (2000) Improved statistical alignment models. Hong Kong, China.

Papadimitriou, C. H., Raghavan, P., Tamaki, H. and Vempala, S. (1998) Latent semantic indexing: A probabilistic analysis. *Proceedings of the 17th ACM Symposium on the Principles of Database Systems*. ACM Press.

Sahlgren, M. (2004) Automatic bilingual lexicon acquisition using random indexing of aligned bilingual data. *Proceedings of the fourth international conference on Language Resources and Evaluation, LREC 2004*.

Sahlgren, M. and Cöster, R. (2004) Using bag-of-concepts to improve the performance of support vector machines in text categorization. *Proceedings of the 20th International Conference on Computational Linguistics, COLING 2004*.

https://doi.org/10.1017/S1351324905003876 Published online by Cambridge University Press