A statistical method for language-independent representation of the topical content of text segments

Thomas K. Landauer and Michael L. Littman
Bell Communications Research
445 South St.
Morristown
NJ, 07960-1910
USA
Tel. 908 829 4255
email=tkl@bellcore.com*

October 24, 1995

Abstract

Where there are texts in more than one language, it would be desirable if users could give queries or examples in the language in which they are most competent and obtain relevant text passages in any language. We have developed and tested a prototype system that makes this possible. The system is based entirely on a statistical technique that requires no humanly constructed dictionary, thesaurus, or term bank. The language-independent representation of text has two steps. In the first, done just once for a subject area, a sample collection of parallel texts—paragraph-by-paragraph translations in two or more languages—is analyzed by a mathematical technique called Singular Value Decomposition. Each word in the sample is assigned a vector value determined by the total pattern of usage of all the words in all the sample paragraphs. In the second step, a new document or query in any of the original languages is assigned a vector value that is an average of the values of the words it contains. Tests on a French-English corpus showed that the method works well.

Key-words: Interlingua, IR, information retrieval, statistical techniques, LSI, SVD, semantics, translation, multilingual filters

1 Introduction

It would be quite useful to have a method for matching text segments in one language with text segments of similar meaning in another language without needing to translate either. Such a facility would obviously be valuable for information retrieval and filtering in circumstances where there are documents or messages in two or more different languages. It could also be used by linguistic and literary researchers, designers of machine translation systems attempting to build interlingual dictionaries and knowledge bases, computational linguists wanting to "align" translated texts, or translators and lexicographers seeking equivalent expressions in alternate languages.

Here we describe an automatic method for reflecting aspects of both the "meaning" of words and the "topic" of text passages in a language-independent manner. The presentation is divided

*A version of this paper appeared in the Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, 1990.

into five parts. In Section 2, we give an informal account of the problem and method. In Section 3, a somewhat more detailed description of the underlying statistical modeling technique is presented, although the reader is referred to other papers for the more mathematical aspects. In Section 4, we describe how the method is actually implemented and applied to language-independent text representation. In Section 5 we report the design and results of an experiment to evaluate how well the representation works in matching texts in French and English. Section 6 offers some conclusions and suggestions for application.

2 Informal description of problem and approach

What we mean by *language-independent* is that words and text are given an abstract description that does not depend on the language in which they were originally produced, and which will be approximately the same for words and passages that mean the same thing but are expressed in different languages. In the method we present, words are assigned numerical codes that represent their pattern of occurrences in a sample of text. Consider a set of paragraphs with both English words and a French translation. Some English word, e_1, occurs twice in the first paragraph, twelve times in the fifth, and so forth. Another English word, e_2, appears once in the tenth paragraph and twice in the twentieth, etc. There will be some French words that appear in similar patterns across paragraphs. In general, the more similar the pattern of a French and and English word, the more similar we would expect their meanings to be. To exploit this reasonable assumption, we need a way to describe patterns of word occurrence abstractly, so that two words that are spelled differently but have the same pattern get the same description. Moreover we want degrees of difference in occurrence patterns to be reflected in differences in the representation. Such a coding scheme would provide the kind of language-independent representation we want.

To understand how such an approach might work, consider the following simple method (that we did not actually try). For every word, A, in one in language, find the word, B, in the other language that most often occurs in the same paragraphs. Assign both words the same arbitrary code number. From now on, code any occurrence of either A or B by this number. The result would be a language-independent representation; and it could be used for new texts in either language. Moreover, if one simply counted how many identical codes any two text passages had in common, passages of similar meaning would usually score higher than entirely unrelated pairs. Of course, this method is much too crude. It does not distinguish between strongly and weakly related A-B pairs, it counts each A as related to only one B, it ignores relations among the A words, etc. With a deeper statistical analysis we can do much better.

The technique we actually used was an extension of *Latent Semantic Indexing* [1, 3, 2]. In this method, a powerful mathematical computation is used to analyze and represent all the linear dependencies between word occurrences in paragraphs (or other text segments). For the two language case, this means that it tries to find the best way to use all the A words in a paragraph to predict how many times each known word in language B is used in the translation of the paragraph. To find corresponding text segments in language B, one can simply rank potential targets by how well they match the prediction. (In actuality, the technique is somewhat more complicated than this in order to address certain issues in statistical estimation, representational economy, and lexical synonymy. More details are given or referenced below.)

3 The Latent Semantic Indexing method

In practice, the method works as follows. An initial sample of multi-lingual documents is submitted to a statistical analysis in order to *train* an automatic interlingual indexing system, i.e. to derive a common representation scheme for all the words in the multilingual sample. After this one-time training, the system can represent any new document, passage or query given in any one or more of the original languages according to a set of derived indexing variables that are language-independent.

The rest of this section and the next give a brief outline of the mathematical basis for the technique. For readers who are not interested in these technical matters, the following may suffice. After the analysis of the training sample, each word is represented by a set of numbers. For example, the word "small" might be represented as $(.2, .9, .4)$, "petite" as $(.2, .9, .3)$, "house" as $(.6, .7, .6)$, and "chambre" as $(.6, .3, .7)$. The similarity of any two sets of numbers reflects similarity in patterns of use of the words in the training paragraphs. In new English-only text, the phrase "small house" would be coded as the average $(.4, .8, .5)$. The French phrase "petite chambre" would be $(.4, .6, .5)$. As you can see, the English and French phrases would get similar code patterns reflecting the similarity of the words used in them. Much expanded, this is just how the system works. The trick, of course, is in how to assign the sets of numbers to words in the best way. Fortunately, there is a mathematical procedure that will insure a particular kind of success. In essence, it chooses the numbers so that the average for the English words will be the same as the average for the French words for every paragraph in the training sample and so that the amount of difference between any two words reflects the difference in their patterns of occurrence. (If you think that's easy, try doing it by hand! It takes a very large amount of powerful and sophisticated computation.) The proof that this representation captures the right kind of similarity is that new paragraphs represented in this way match paragraphs in other languages that have the same meaning. Section 5 demonstrates this kind of success for a large sample of expertly translated paragraphs.

We now give a more technical introduction to the method. To follow this, you will need some knowledge of linear or matrix algebra. The Latent Semantic Indexing (LSI) approach uses matrix decomposition procedures to model the underlying correlational structure of the distribution of terms in documents.

The modeling is accomplished by applying the factorization method Singular Value Decomposition to a matrix formed from all the words (the rows) occurring in all the original text segments (the columns); each cell of the matrix contains the number of times that a particular word appeared in a particular text segment. In the result, the original $T \times D$ (term-by-document) matrix is approximated by the product of three lower rank matrices of orthogonal derived indexing variables. The first and third matrix represent terms and documents respectively as values on a smaller set of independent "basis" vectors; the second matrix contains scaling coefficients. Mathematics and computational implementation are presented in detail in Deerwester, Dumais, Furnas, Landauer and Harshman [1].

The retrieval process is the same as in certain standard vector information-retrieval or pattern recognition methods, e.g. using the cosine between a document and query vector or between two document or term vectors as the measure of similarity. The principal difference between LSI and previous vector models (See Harman [4], Salton and McGill [6]) that have been used in information retrieval is that the vectors are constructed in a space with many fewer dimensions (typically 100) than the number of original terms (typically many thousand), and that these dimensions are the subset of linearly independent basis vectors by which the original term-by-document matrix can best be approximated in a least squares sense.

The dimension reduction step of LSI has the advantageous property that small sources of variability in term usage are dropped and only the most important sources kept. Among other things, the method causes synonyms or near synonyms to be represented by the same or similar vectors. As a result, queries can automatically match similar text, e.g. translations, with which they share no terms. This cannot be accomplished with the usual representations used in information retrieval without manually constructed thesauri with their attendant expense and conceptual difficulties [5]. (More traditional natural language understanding systems require complex term-knowledge bases and extremely labor-intensive information entry to represent lexical relations for this purpose, and are therefore not easily applied to new domains with large vocabularies.)

For the cross-language matching situation, it is especially important to understand the synonym equivalencing effect of LSI. Consider an idealized case of translation in which each of the n terms in one language is distributed independently from any other term in that language but has exactly one corresponding term in the other language. Corresponding terms would, ipso facto, be distributed in the same way across documents. Thus the term-by-document matrix of a set of translated documents, each document represented as the union of its two different language versions, would contain just n pairs of identical term rows. An SVD analysis would disclose that the full $T \times D$ matrix could be perfectly reconstructed with just n instead of the original $2n$ dimensions. In the derived space, the vectors representing corresponding words in the two language would be identical. In other words, if only the n best dimensions were retained, the same representation would be obtained for documents described in either language. Of course, any real translations will not conform to these ideal assumptions, but they may approximate them well enough to make cross-language document matches effective.

(Note that while the SVD model used in this way bears some resemblance to clustering methods that have previously been applied to word representation with limited success, it is also different in important ways. In particular, where clustering results in a finite number of sets within which all terms are treated as identical, SVD produces a representation in which the similarity between any two terms is a continuous variable with a potentially unique value. In other words, LSI provides a model of the similarity structure in lexical choice that can extract many more parameter values, and much more information, from the data than does clustering.)

The LSI method has previously been applied only within a single language. The number of dimensions retained has been determined by empirical trial and error; optimal matching performance has usually been obtained with about 100 dimensions for collections of many hundreds to several thousands of documents. The synonym equivalencing property has been observed to occur often in practice, and overall recall/precision performance for information retrieval in standard test collections has varied from a few percent less good to almost 30 percent better for LSI than for the best previous methods, i.e. standard vector methods with term weighting [6].

4 Cross-Language use of LSI

To apply LSI to cross-language retrieval, we first analyze a *training set* of paragraphs for which translations exist in two (or more) languages. The term-by-document matrix of this set is composed from the union of the terms in their renditions in the two (or more) languages. This matrix is analyzed by singular value decomposition. The resulting representation defines vectors for terms in both languages, as well as for the original set of paragraphs. Thus, this representation allows the initial set of paragraphs to be accessed by queries in either language alone. More importantly, once the training analysis has been completed, new texts can be given language independent representations on the basis of terms from any one language alone. In the derived indexing space there is a point representing each term in the training set. A new single-language text segment is assigned a point in the same space by putting it at an appropriate average of the location of all the terms it contains. For cross-language matching, the same number of dimensions are kept as would be required to represent the sample in a single language. The dimension reduction step insures that terms that occur in similar contexts (i.e. that have similar meaning) will be given similar values (even if they never occur in precisely the same paragraphs), and, in addition, reduces storage requirements, and presumably improves the estimation of term-term associations by a kind of "smoothing" effect.

5 Evaluation experiment

To test how well the method generates the same representation for text of ostensibly the same meaning in different languages, we applied it to a sample of parallel text in French and English. The sample corpus was taken from the so-called "Hansard" publication, the Proceedings of the Canadian Parliament. These are speeches and other entries on a variety of topics, in both the original version and liberal translation on a paragraph by paragraph basis. The initial SVD analysis was performed on 900 randomly chosen dual-language paragraphs, i.e. in both English and French. This generated a 100 dimensional language-independent indexing space and the corresponding vectors for over 3,000 words in each of the two languages. Next, another 1,582 randomly chosen paragraphs, none of which were used in deriving the indexing space, were coded as the 100-vector average of their term vectors, using only French words from their French versions. We then computed a vector code for each English paragraph, on the basis of only its English words, and compared it to each of the coded French paragraphs. (Comparisons were computed as cosines between the two vectors.) Among the 1,582 French paragraphs, the one most similar to a given English paragraph was its French translation 92% of the the time. In other words, the matching was perfect more than nine times out of ten.

(The average cosine between French and English folded-in versions of the same paragraph was 0.78, s.d.=.09. By comparison, cosines between 200 randomly sampled pairs of *different* French and English paragraphs, i.e. ones that were not translations of each other, averaged 0.21, s.d.= .10.)

To gain some insight into what would happen in a typical information retrieval application, where short and imprecise queries are the norm, we studied the consistency with which a short piece of text and its translation matched the same paragraphs. We selected a set of 20 individual English sentences taken from other parts of the data set as *queries*. We compared these with the test paragraphs indexed by only one language, and for each returned the closest 10 paragraphs. That is, we used coded English queries to search for both French-only and English-only documents that had been coded without translation. We repeated this procedure starting with the 20 French translations of the English queries. On the average 4.1 of the ten paragraphs most similar to a sentence were identical for the French and English versions of the *query* sentence. As a control, we compared how consistently French and English versions of a query returned the same paragraphs when the documents were also translated. To do so, we repeated the training analysis using the same 900 training paragraphs, but separately in their English and French renditions, i.e. there were separate representations for French and English rather than a single language-independent representation for both. We then coded the remaining paragraphs into an English-based space on the basis of their English versions (original or translated as appropriate) and into a French-based space on the basis of their French versions. Again we retrieved the 10 most similar paragraphs for each sentence, in this case separately for French sentences against French paragraphs and English sentences against English paragraphs. An average of 3.1 of the ten were the same for the French and English versions of the *query* sentences. The differences in averages, 4.1 vs 3.1, is statistically reliable (sign test, $p < .03$). This implies that the cross-language LSI technique, in which queries are entered in their presented language and all comparisons are made in the language-independent representation, would yield greater consistency across differences in query languages than would a comparable method in which queries were manually translated into the original language of the target paragraphs.

Another way to gain some intuition as to the performance of the method is to examine correspondences between terms in the derived space. To do this, we chose English words at random and searched for the closest French words. A sample of 40 English words are presented in Table 1, along with two to four of their closest French word neighbors, (that is those whose vectors form relatively small angles in the 100-dimensional space). It appears that very close neighbors, words with cosines above about 0.8 are very often either direct translations or words that are highly related semantically or collocationally in the current context (e.g. languages/langues(.98), questioning/legitime(.89)). That is, close neighbors in the latent semantic space include not only pairs that are synonyms in the usual sense, but also words of related meanings that tend to occur in similar contexts in the domain of discourse. Note, for example, that the English "house" is represented as very close to to the French "chambre" in this sample of Parliamentary discourse. Word pairs with moderately high similarities, between .5 and .8, contain a mixture of literal translations, what we would call "interesting companions," and incomprehensible associations (e.g. very/tres(.72), welfare/cigarettes(.67), demand/meres(.77)). When a French neighbor of an English term has an angle to it with cosine less than .50, the relation is usually uninterpretable and apparently largely due to chance (.e.g. things/ridings (.45)), , but not always, (e.g. it/il(.46)). (Note that due to software limitations the French words were analyzed, and therefore are presented here, without accent marks.)

6 Some implications, applications and extensions

First, let us note that it is not logically necessary, and may indeed be surprising, that term vector methods like the present one perform well for matching the meaning or topic of text. In such methods text segments are represented, in essence, only as "bags of words," histograms of contained-word frequencies. All information that is transmitted by word order (i.e. that would require syntactic parsing to recover) is ignored. Nevertheless, when a paragraph was treated as a query in the cross-language tests, the most similar paragraph in nearly every case was the translation of the same text. Apparently, then, textual meaning, at least up to the level of expressive precision needed to select one paragraph from among 1,582 from a similar source, is carried in the combination of words used, independent of syntax, argument order and the like. (The insensitivity to liberal translation demonstrated here indicates that it is the choice of a set of words as semantic units to convey meaning, not just the accidental profile of lexical terms, that results in selectivity.) While the considerable representational power of unordered collections of words may be a depressing observation from a humanistic, literary, or even a traditional linguistic or "natural language understanding" point of view, it should be viewed as encouraging in terms of the more general goal of practical automatic language processing.

As for extensions of the method, we will mention only some possible uses as an aid for translation. While fully automatic translation involves hard problems of syntax, language generation, pragmatics, etc. to which LSI provides no direct solution, there are at least three ways in which it might serve as a useful adjunct to manual or automatic translation.

First, a very interesting use for LSI in translation would be for automatic lexical disambiguation.

it que 0.56 est 0.53 il 0.46	*changes* changements 0.83 outillage 0.60 desormais 0.60	*use* utilisent 0.62 utiliser 0.62 lignes 0.52	*that* que 0.72 le 0.61 il 0.57
our nos 0.76 notre 0.76 confident 0.59	*was* etait 0.68 avait 0.59 ete 0.46	*talk* belles 0.58 sexe 0.58 celebrer 0.57	*companies* petrolieres 0.72 societes 0.70 carte 0.66
not pas 0.86 ne 0.80 que 0.43	*way* facon 0.60 maniere 0.39 sorte 0.37	*of* de 0.81 la 0.77 le 0.72	*things* politiquement 0.46 ridings 0.45
unique unique 1.00 desobligeantes 0.66 desobligeants 0.64	*this* cette 0.58 ce 0.47 de 0.45	*statement* facteur 0.44 petro 0.43	*house* chambre 0.91 communes 0.55 greffier 0.45
member depute 0.93 interrompre 0.54 demanderais 0.51	*debate* debat 0.97 partis 0.58 discuter 0.52	*unless* petroliers 0.76 frappe 0.75 lourdement 0.72	*in* de 0.69 en 0.66 le 0.64 dans 0.60
prior presidence 0.58 parole 0.56 pourrions 0.54	*clear* clairement 0.57 lacunes 0.52 reponse 0.51	*welfare* interdire 0.67 cigarettes 0.67 finlandais 0.66	*demand* meres 0.77 lacunes 0.66 precedentes 0.64
languages langues 0.98 langue 0.78 patrimoine 0.74 multiculturel 0.72	*through* appui 0.50 entremise 0.39	*very* tres 0.72 graves 0.45 surement 0.38	*being* caractere 0.49 soumis 0.40
they ils 0.86 ont 0.48 disent 0.42 leur 0.41	*question* adresse 0.64 ma 0.60 algoma 0.47 poser 0.46	*financial* financier 0.88 financieres 0.86 aborde 0.71	*questioning* legitime 0.89 normales 0.84 cotret 0.82
occasion exprimer 0.60 profite 0.56 anniversaire 0.53	*related* causees 0.61 connexes 0.57	*themselves* assainissement 0.73 demenager 0.72 laissee 0.69	*help* aider 0.66 ecole 0.62 aide 0.51 besoin 0.50
share partage 0.63 generale 0.50	*business* genres 0.63 lancer 0.62 planification 0.61	*world* monde 0.67 etions 0.63	*bill* projet 0.95 loi 0.92

Table 1: English Word with Nearest French Neighbors

When a word in one language translates by dictionary to two or more words in another, the location of the alternative words in the LSI space relative to the location of the surrounding text could be used to make a choice. That is, when an ambiguity is discovered, vectors for the remaining words of the phrase, sentence, paragraph, etc., depending on the amount of context appropriate, would be compared with each alternative, and the sense with the best match selected.

(Applications to sense disambiguation within a single language can also be imagined. If text sources in which senses were marked were available for analysis, separate LSI vector representations could be derived and subsequently matched against vector representations of sentential or discourse context.)

Second, as illustrated in Table 1, the method can produce lists of words from two or more languages at least some of which are approximate equivalents or expressive companions in a particular restricted domain of discourse or a particular corpus. Translators, either human or automatic, might find such lists useful adjuncts to traditional dictionaries and thesauri. Note that the similarities between words derived from LSI are not simple first-order co-occurrence coefficients. Because the dimension-reduced SVD model computes the best linear estimates of document-to-term correspondence using all the data in the matrix, the degree of similarity between the vectors of two terms depends on indirect as well as direct associations. Just as two documents that share no terms can be placed in nearby positions, terms that never occur in the same document can be estimated to be similar if the documents in which they do occur are similar. Less formally, words with nearby vectors will often be ones that authors might well have used in the same documents, even if they never actually did so in the corpus at hand. Estimated similarities can also be smaller than simple correlation would suggest, if, for example, the words in question occur in documents that are widely separated in the derived space.

A third manner of aiding translation would be simply to provide manual translators with the LSI-based information retrieval system to run over the text base in which they operate. This would allow them to locate documents, paragraphs, or perhaps sentences, on similar topics or containing similar words in any language in order to study appropriate expressions and terminology. Such a tool might also be useful in lexicography for finding related expressions.

7 Summary

The Latent Semantic Indexing technique was applied to cross-language text matching. An initial training set of text (paragraphs from Canadian Parliamentary debates) existing in two languages was first analyzed to construct a derived indexing space defined by an optimal set of 100 orthogonal factors. New paragraphs were then coded as the average of their single-language term vectors. It was found that paragraphs were closer to their translations than to any other different-language paragraph more than nine times in ten.

We also observed that when queries formed from short unrelated sentences in the text base were used to find their closest paragraphs in each language, the same nearby documents tended to be returned for a query and its translation more often with this method than in a control condition in which documents were searched in the same language as the query. In addition, we illustrated the ability of the method to suggest cross-language lexical equivalents based on overall usage patterns in a corpus.

The experiments indicated that the method can be successfully employed when only a modest amount of translated training text is available; in the present case 900 paragraphs were sufficient.

We suggest that the technique could be very useful for certain large multilingual document collections or text sources, and as an adjunct to translation tools or systems.

Acknowledgements

We thank all the participants in the LSI research effort: George Furnas, Richard Harshman, Susan Dumais, Lynn Streeter, Karen Lochbaum, Scott Deerwester, and Laura Beck for providing the foundation for this application. We thank Jerry Proulx of Parliamentary Publications, Canada, for efficient and generous provision of the bilingual data, and Michael Lesk for help in its pre-processing. We thank specially Susan Dumais, Scott Deerwester and George Furnas for practical implementation help, theoretical discussion, and expository advice.

References

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[2] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings CHI '88*, pages 281–286, 1988.

[3] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a Singular Value Decomposition model of latent semantic structure. In *Proceedings of the 11th ACM International Conference on Research and Development in Information Retrieval*, 1988.

[4] D. Harman. An experimental study of factors important in document ranking. In *Proceedings of 9th ACM Conference on Research and Development in Information Retrieval*, 1986.

[5] G. Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Sciences*, 21:187–194, 1970.

[6] G. Salton and M. J. McGill. *Automatic information organization and retrieval.* McGraw-Hill, New York, 1983.