An LSA Implementation Against Parallel Texts in French and English

Katri A. Clodfelder
Indiana University
kclodfel@indiana.edu

Abstract

This paper presents the results of applying the Latent Semantic Analysis (LSA) methodology to a small collection of parallel texts in French and English. The goal of the analysis was to determine what the methodology might reveal regarding the difficulty level of either the machine-translation (MT) task or the text-alignment (TA) task.

In a perfectly parallel corpus where the texts are exactly aligned, it is expected that the word distributions between the two languages be perfectly symmetrical. Where they are symmetrical, the difficulty level of the machine-translation or the text-alignment task should be low. The results of this analysis show that even in a perfectly aligned corpus, the word distributions between the two languages deviate and because they do, LSA may contribute much to our understanding of the difficulty of the MT and TA tasks.

1. Credits

This paper discusses an implementation of the Latent Semantic Analysis (LSA) methodology against a small collection of perfectly parallel texts in French and English[1]. The texts were made available by the HLT-NAACL and are taken from daily House journals of the Canadian Parliament. They were edited by Ulrich Germann. The LSA procedures were implemented in R, a system for statistical computation and graphics, and were written by John C. Paolillo at Indiana University, Bloomington.

2. Introduction

LSA is an analytical methodology that uses mathematical procedures and vector space modeling techniques to generate an abstract, numerical representation of the relationships among words and documents in a collection of texts (the corpus). In this analysis, the methodology is used to identify the symmetry that exists among the pattern of relationships and associations in the parallel texts. Where texts are perfectly aligned, it is expected that for every occurrence of a word in one language, an exact correspondent exists in the other language. However, the analysis shows that even in a perfectly aligned corpus, the word distributions between the two languages deviate and a one-to-one association does not exist.

An example of how word symmetry breaks down in parallel texts can be seen in two "sets" of parallel documents, F1-E1 and F2-E2. In these paired documents, the cross-language term correspondence between the French term "je" and the English term "I" in the two sets shows that in the first pair, "je" occurs 42 times and "I" occurs only 37 times. In the second pair, "je" occurs 59 times in the French document and 62 times in the English document. Such differences in word usage patterns between corresponding terms are very common and create difficulties for the MT or TA tasks.

Because of the way LSA represents word-usage associations and patterns among documents and terms, it may have much to offer in understanding the difficulty levels of these tasks. This analysis shows that, in spite of usage differences resulting in non-symmetrical cross-language word distributions between the corresponding terms of any given language pair, the LSA methodology is capable of identifying the appropriate usage pattern for each of the terms, within its own language. This paper presents a first look at the alignment patterns found in a parallel corpus and how LSA may offer some insights into the MA and TA tasks.

3. LSA

The LSA methodology begins with the term-by-document matrix, an *n x m matrix* where each value in the matrix is the frequecy of the *nth* word in the *mth* document. A weighting procedure is applied that weights each of the term frequencies (TF) by the

[1] At present, the collection of texts used consists of 30 English and 30 French documents, all perfectly mated. Cognates were not distinguished between languages, e.g., revolution is counted as both a French and an English term.

inverse document frequency (IDF)[2] (Salton, G. et al 1968). A very powerful mathematical procedure, known as singular value decomposition (SVD), is then performed against the transformed matrix. SVD permits the reduction of any *n x m matrix* to a set of three matrices, such that $M = U\Sigma V^T$, where

U = (*m x m matrix of left singular vectors*)
Σ = (*n x m diagonal matrix containing the singular values of M*)
V^T = (*n x n transposed row matrix of the right singular vectors*).

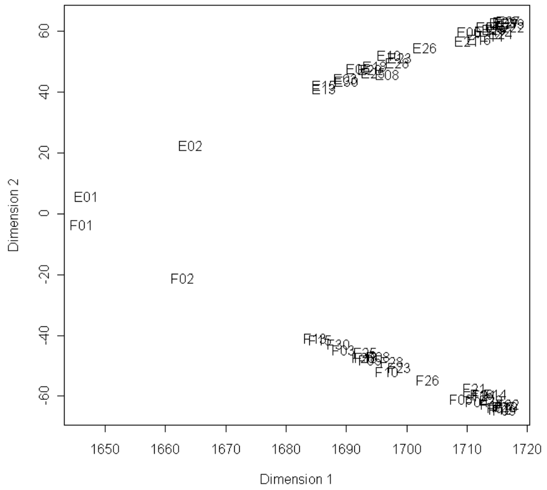
While the SVD solution of any given matrix can re-create the original matrix, exactly, its primary value lies in its capacity to infer what the pattern of relationships and associations is for the words in the documents, if all the linguistic data in the corpus is represented on a smaller number of dimensions than that of the original matrix (Landauer et al 1998).

4. Interpretation

Although it is useful to think of the "dimensions" as representing the document vectors, it should be emphasized that a "dimension" is a more abstract notion related to the contribution that a given document vector makes in explaining the relationships and associations of the linguistic data contained in the corpus. In this section, an example of how the SVD procedure depicts the word and document relationships on the different dimensions is discussed.

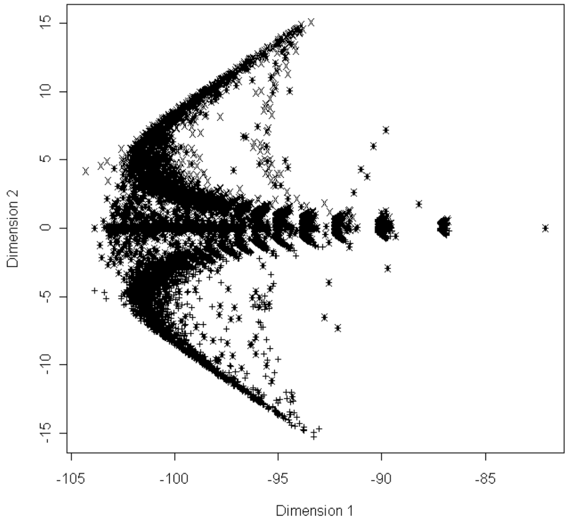
In Figure 1, the location of the documents is shown on Dimensions 1 and 2. Notice that on Dimension 2 (the vertical axis), all of the English documents (E_n) lie above zero and all of the French documents (F_n) lie below. Moreover, the corresponding French and English document pairs are almost perfectly aligned across the horizontal axis. As an abstract notion, Dimension 2 clearly represents the two languages.

On Dimension 1, the documents occur from left to right, in the order of smallest document to largest document, making this dimension representative of document size. No attention was paid to ordering the documents according to size when the data was input into the model. The LSA model is able to identify this relationship, automatically, and represent it, as shown here in Figure 1.



Document Plots on Dimensions 1 and 2
Figure 1

Figure 2 shows the word relationships on Dimensions 1 and 2. The English words are represented by "x" and the French words by "+".



Word Plots on Dimensions 1 and 2
Figure 2

Where the two symbols cross, the data point represents a cognate.[3] In the same way that the documents were split with the English documents occuring above zero

[2] The IDF of term, t, is the ratio of the total number of documents to the total number of documents in which the given term, t, occurs, e.g., if term, *t*, occurred in 6 of the 60 documents in the Corpus, its IDF ratio is 60/6 (or 10).

[3] The term cognate is used to include words such as "chose", which means "thing" in French and is the past tense of "to choose" in English; "pays", which is "country" in French and present tense of "to pay" in English; and so on.

and the French documents below in Figure 1, so are the English and French words split in Figure 2. As shown, the English words are located above zero on the vertical axis and the French words are located below. Whether a cognate is represented as an English term and appears above zero or as a French term appearing below, is entirely dependent on which set of language documents "drives" the association for the cognate.

Figure 2 shows that, in spite of the very high degree of symmetry between the document-pairs of the French and English texts (Figure 1), the cross-language patterns of association among the words in the documents are not completely symmetrical. If they were, there would be a corresponding "x" sign above the horizontal axis in exactly the same position as every "+" sign below the axis. From an MT or TA perspective, the greater the degree of cross-language symmetry among words in the documents, the easier the task of selecting the appropriate target term. When cross-language symmetry is low, however, the task of finding the appropriate target term is more difficult.

5. Symmetry of Query Results

For the most part, single-term queries accurately identified the most relevant, same-language documents and they did so in spite of non-symmetrical, language-specific, usage-associations of the query terms. For example, in Table 1, the query using the English term "aboriginals" returned E22 as its most relevant document; however, the query using the corresponding French term "autochtones" returned F09 as its most relevant document.

Query E	Query F
ABORIGINALS	AUTOCHTONES
E22	F09
E09	F17
E17	F07
E11	F06
E27	F27
E19	F19
E12	F12
E07	F11
E24	F04
E29	F23

Aboriginal-Autochtones Query Results
Table 1

At first glance, these query results would seem to be undesirable. However, they are perfectly consistent with the language-specific usage patterns of these two terms.

In French, because of number agreement between adjectives and nouns, the plural form of "autochtone" is used quite frequently in comparison to the plural form of its English counterpart "aboriginal", where number agreement is not required. For example, the French usage of "les (peuples) autochtones" is often realized as "the aboriginal people(s)" in the corresponding English document. The impact of this non-symmetrical usage pattern of corresponding language terms is seen in the query results. While the most relevant French document, F09, contains 50 occurrences of the plural "autochtones", its corresponding English document (returned as second relevant in Query E) contains two occurrences of the plural "aboriginals" and 49 occurrences of the singular "aboriginal".

Continuing on with this example, Query E identified the English document, E22, as the most relevant to the query term. E22 contains nine occurrences of the plural "aboriginals" and no occurrences of the singular "aboriginal". Thus, the results of Queries E and F shown in Table 1 demonstrate that not only is the LSA methodology sensitive to the language-specific word distributions of cross-language word pairs, it is also capable of distinguishing those distributional variations in order to identify the most relevant documents for the language of the query term, accordingly. In other words, given the dissimilarity in the distributions of the plural forms of each term in the cross-language word pair, the LSA methodology behaved appropriately for each of the queries.

The order of the documents returned as relevant to the queries in Table 1 is important from an MT and TA perspective also because it shows that LSA has some capability to "align" similar, but not exact, terms. For example, in Query E, the first three documents all contain the exact query term, "aboriginals". The next five documents contain only the singular form of the query term. The remaining two documents do not contain either form of the query term. In other words, after the documents that contained the exact query term, the LSA methodology chose as more relevant, documents containing the singular form of the query term over documents that contained no form of the query term. If the LSA methodology were only identifying relevant documents on the basis of finding terms with an exact match to the query term, it would have no preference for choosing documents containing the singular form of the query term over documents that contained no form of the query term.

6. Conclusion

This paper presented a brief discussion on the possibility that the LSA methodology may have something to contribute with respect to identifying the difficulty level of the MT and TA tasks. In particular, it was shown that LSA can represent the symmetrical and non-symmetrical relationships that exist among the terms in cross-language document pairs. It was also shown that LSA has some capability to "align" similar terms in order to identify relevant documents in response to a query.

Such positive results using the large, non-homegenous documents that were used in this analysis are very promising and suggest that further research in this area is needed. Additional work is planned that will separate the documents into smaller, syntactical units that will be analyzed using a very similar approach. However, the words and "documents" represented in the semantic space that is generated by these very small syntactic units will have very different associations and relationships than they do as part of the larger, non-homegenous texts.

It is believed that by restricting the input texts to very small syntactic units, the LSA methodology will be able to make the proper "alignment" associations between the cross-language word pairs and, as a result, provide some information regarding the types of word pairs that are most difficult to align by an MT or TA system.

Finally, the small number of documents used in this analysis[4] raises questions about the optimal number of input texts that are needed to obtain valid results using the LSA methodology. It may be that a much smaller number of input texts is needed than formerly believed, in order to "train" an LSA-based system to correlate the appropriate cross-language word pairs.

References

Laundauer, Foltz, Laham. 1998. "An Introduction to Latent Semantic Analysis." Discourse Processses, 25, 259-284. 1998.

Rehder, Bob, Michael L. Littman, Susan Dumais, Thomas K. Landauer. 1997. "Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing." TREC-6, 233-239.

Salton, G., and Lesk, M. E. 1968. "Computer evaluation of indexing and text processing." Journal of the Association for Computing Machinery 15.1.

[4] 30 mated pairs (or 60 documents in total)