International Journal of Linguistics, Literature and Translation (IJLLT)
ISSN: 2617-0299
www.ijllt.org

Latent Semantic Analysis and Machine Translation

Dr. Mahmoud Mobaraki [1]* & Dr. Abolfazl Mosaffa Jahromi [2]
[1]*Assistant Professor, Linguistic Department, Faculty of Humanities , Jahrom University, Iran, Jahrom*
[2] *Assistant Professor, Linguistic Department, Faculty of Humanities, Jahrom University, Iran, Jahrom*
Corresponding Author: Dr. Mahmoud Mobaraki, E-mail: mmobaraki@jahromu.ac.ir

ARTICLE INFO

Received: October 07, 2018
Accepted: October 20, 2018
Published: November 30, 2018
Volume: 1
Issue: 4
DOI: 10.32996/ijllt.2018.1.4.5

KEYWORDS

computational linguistics, latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA), machine translation, coherence, irony

ABSTRACT

Computer-based translation systems are not rivals to human translators, but they are aids to enable them to increase productivity in technical translation. Machine translation aims to undertake the whole translation process, but whose input must invariably be revised. Latent Semantic Analysis and Probabilistic Latent Semantic Analysis are two newly developed computational models which their application in machine translation will solve some of the problems facing machines in accounting for the way human knowledge is comprehended.

INTRODUCTION

The field of machine translation (MT) has been the pioneer research area in computational linguistics during the 1950s and 1960s. When it began, the goal was the automatic translation of all kinds of texts at the quality of human translator. It became very soon apparent that this goal was impossible. However, it was found that for many purposes MT output could be useful to those who wanted to get a general idea of the content of a text in an unknown language. But machine translation was constrained by limitations of hardware, in particular by inadequate computer memories and slow access to storage of dictionaries and text, and by the unavailability of high-level programming languages. Syntax was a relatively neglected area of linguistic study and semantics was virtually ignored. The researchers knew that whatever system they could develop would produce poor quality results. In this atmosphere, the translations produced were impressively colloquial, based on small vocabularies and carefully selected texts (Jan, 2001).

In the next decade, by improved computer hardware, especially developments in syntactic analysis based on research in formal grammars (e.g. by Chomsky), it was assumed that the goal of MT must be the development of fully automatic systems producing high quality translation. The emphasis of research was therefore on the search for theories and methods for the achievement of perfect translation (ibid.). The idea of "fully automatic high quality translation" was criticized by Bar-Hillel (1960) and progress in this area proved no fully automatic system capable of good quality translation. The systems produced poor translated texts and as a means of improving the quality vocabulary, structure and style of the text before input to the systems were controlled. But the output produced needed to be edited, and now still the inevitably imperfect nature of MT output is stressed.

During next decades from 1970s, there has gradually been some improvement of translation quality, although not as rapidly as many would have hoped (Hutchins, 1986 & 1988). In general, improvement in this field came from research building upon computational and linguistic methods and techniques.

Machine translation is, therefore, under the influence of new linguistic and computational techniques and the principal focus of MT research remains the development of systems for translating scientific documents and other texts whose style is not important part of the message. Machine translation initially used dictionary based approach, i.e. word-for-word translation and the use of statistical method was advocated by Warren Wear in 1949. Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA) are two newly developed computational techniques which are applicable in MT. The strength of these two techniques lies in their independence of any language structure and being able to account for the way knowledge is being used in contexts by humans. This article tries to introduce LSA and PLSA and at the same time provide an overlook into the use of these techniques in MT, their advantages and drawbacks in solving some problems of MT like irony, metaphor, polysemy, coherence and topic shift.

LITERATURE REVIEW

Latent Semantic Analysis (LSA, also known as Latent Semantic Indexing, or LSI) is a well-developed technique for representing word and passage meanings as vectors in a high dimensional "semantic" space. Through application of linear algebra methods singular value decomposition and dimensional reduction, a co-occurrence matrix is transformed to better reflect the "latent," or hidden, similarities between words and documents. The technique can be used to determine the most likely meaning of a polysemous word from some given context by comparing a vector constructed from that context with document vectors. Vectors representing similar passage meanings should be near each other, as LSA is said by some of its creators to "closely approximate human judgments of meaning similarity between words" (Landauer and et al. 1998).

Most studies to date have focused on LSA's applications in searching and document retrieval. In this field, LSA has been shown to offer a marked improvement over other methods (Dumais, 1994). Cross-language information retrieval search results in languages differing from the query has also received attention (Rehder and et al. 1998) as has LSA's use in language modeling (Kim and Khudanpur, 2004).
LSA has also been tried with human vocabulary synonym and word-sorting tests, in the course of research on how well LSA models human conceptual knowledge, and scored not far below group norms (Landauer and et al. 1998). On the practical side, LSA has been used in a commercial product called the "Intelligent Essay Assessor," which evaluates students' knowledge and writing skills (Landauer and et al. 2000).

However, at least one study has addressed LSA's potential in machine translation, specifically in dealing with polysemy in Korean-English translation (Kim and et al. 2002). This study did not use the general context of an ambiguous word, but rather considered a single argument word in a specific grammatical relationship, such as subject-verb, between the argument and the target polysemous word. The correct meaning of the target was drawn

from a dictionary storing examples of argument words. If the given argument did not appear in the dictionary, the correct translation class was that of the example word most similar to the argument. The project used an LSA model to determine this similarity by finding the example word whose vector representation was closest to the argument word's, under the theory that words of similar meaning are "close" in the semantic space. Thus, this LSA model relied on vector representations of individual argument and example words rather than on representations more closely associated to the meaning of the polysemous words themselves.

DISCUSSION

Translation has been defined as the production of a text in TL with the same effect in SL (Newmark, 1981). Part of producing the same effect in TL is to know how words are perceived and comprehended in TL different contexts. The question is that how it is possible to account for different contextual usage of words in translation. How it is possible to know which word is most probable to occur in a given context? Comprehension of the text in the target language based on its contextual usage is the key point which plays a fundamental role in depicting the way language is processed in TL. In this way, the texts produced in translation will be perceived and comprehended more naturally because the texts will be comprehended as it is stored in the mind and retrieved in different contexts. The aim this article seeks is to prepare for the way of facilitating the translation especially machine translation by relying on contextual usage of words in TL. Latent semantic analysis (LSA) is the framework used to give the solution to some problems facing computer to cope with.

Latent semantic analysis and translation

Latent semantic analysis is a general theory of acquired similarity and knowledge representation. It ignores all linguistic structures in the text including syntax, morphology, etc, and is sensitive only to occurrences of words. The basic assumption of LSA is that the words which have similar meanings tend to occur in similar contexts. LSA's power lies in the fact that it is sensitive not only to direct co-occurrences, but can also infer indirect relations between words across texts. Measuring LSA in translation will enable machine to cope with some drawbacks that face machine in choosing between words while translating into another language. This model is able to represent complex semantic structures of given contexts in TL. This fact will help to provide the reader the structures above the structure of language which produces the same effect as SL. This model does not require any human-like knowledge in translation which enables the machine to perform the task of translation as efficiently as possible without relying on intelligence or world view.

LSA extracts and infers relations of expected contextual usage of words in passages of discourse. It uses no human-made dictionaries, syntactic parser or the like. Only raw text parsed into unique character strings is used as input data.

Next, LSA applies singular value decomposition (SVD) to the matrix. SVD is a form of factor analysis and defined as

$$A = U \sum V^T$$

where $\sum$ is a diagonal matrix composed of nonzero eigenvalues of AA^t or A^tA and U and V are the orthogonal eigenvectors associated with the r nonzero eigenvalues of AA^t or A^tA respectively(Kim, Chang, Tak Zhang, 2002)

LSA is a valuable analysis tool with wide range of applications
(Deerwester, Dumais and Landauer, 1990; Foltz and Dumais, 1992; Landauer and Dumais, 1997). Application of LSA in machine translation will improve its efficiency beyond that of translation done without LSA at hand.

More on LSA and translation

Latent semantic analysis (LSA) is a theory and method for extracting and representing the contextual meaning of words by statistical computations applied to a large corpus of text (Landauer & Dumais, 1997). The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other (Landauer, Foltz & Laham, 1998). By the application of LSA in translation it is possible to predict automatically whether a word can occur or not based on its frequency of occurrence and its correlation with other words especially the topic of a given context. LSA exploits a new theory of knowledge induction and representation (Landauer & Dumais, 1997, 1996) that provides a method for determining the similarity of meaning of words and passages by analysis of large text corpora. Translation by the use of LSA can account for contextual use of words as they are produced in different contexts of TL. This point has the advantage of ignoring what the collocation or usage of SL words may be. Another advantage is that it enables the machine to make decision beyond the structure of language.

LSA constitutes a fundamental computational theory of representation. Its underlying mechanism can account for a long-standing and important mystery: the inductive property of learning by which people acquire much more knowledge than appears to be available in experience, the infamous problem of the "insufficiency of evidence" or "poverty of input". The role of LSA in machine translation will be capturing information contained in contextual usage of words in relation to experience i.e. the knowledge that machine falls foul of in translation. The inductive nature of this method inculcates indirectly the way knowledge is imparted in human cognition and by invoking LSA in translation from SL into TL by relying on TL experience is come up with based on the way it is encoded there.

LSA is a fully automatic mathematical and statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. It is not a traditional natural language processing or artificial intelligence program. It uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, morphologies or the like, and it takes as its input only raw text parsed into words and separated into meaningful passages or samples such as sentences or paragraphs (Landuer et al, 1998). Because no information other than contextual usage based on mathematical computation plays role in LSA, it can properly be used both to SL and TL without restriction.

LSA estimates the frequency of occurrence of words in different contexts and based on working out the correlation between two words, it can predict whether a word can co-occur with another word in a given context or not. A machine will be able to predict which word in TL has the most correlation with which word or words. In this way, LSA enables machine in making choice and to organize the text as it is imparted in human cognition and reflected in text-types. By accounting for LSA in SL and TL it is possible to produce more natural human-like translation. In this fashion, a machine can simulate human knowledge in translation by working out the usage of words in different contexts without further linguistic prior knowledge.

LSA by accounting for contextual usage of words in SL and TL will enable the machine to translate based on contextual usage of TL ignoring SL, hence producing text based on the way knowledge is perceived by human in TL. This is the crucial point in producing more natural text. Note that much of the information that LSA uses to infer relation among words is in data about passages in which particular words does not occur. LSA can be used to determine the coherence of texts (Landauer and Dumais, 1997; Foltz, kintsch and Ladauer, 1998). The result of the analysis of the Britton and Gulgoz (1991) and McNamara et al (1996) indicates that LSA can provide an accurate model coherence of the text. LSA provides a fully automatic method for comparing units of textual information to each other to determine their semantic relatedness. These units of text are compared to each other using a derived measure of their similarity of meaning. This measure is based on a powerful mathematical analysis of direct and indirect relation among words in a large corpus. Semantic relatedness corresponds to a measure of coherence because it captures the extent to which two text units have semantically related information. By LSA in hand machine translation is able to account for coherence without relying on counting literal word overlap between units of text. LSA's comparisons are based on a derived semantic relatedness measurement that reflects semantic similarity among synonyms, antonyms, hyponyms, compounds and other words that tend to be used in similar contexts. As the power of computing semantic relatedness with LSA comes from analyzing a large number of text examples, for computing the coherence of a target text in translation, it may first be necessary to have another set of texts that contain a large proportion of the terms used in the target text and that have occurrences in many contexts. One approach is to use a large number of encyclopedia articles on similar topics as the target text in translation. With accounting for coherence translation based on TL, LSA provides a reader a well-connected representation of the information in TL. This connected representation is based on linking related pieces of textual information that occur throughout the text. The linking of information in translation by application of LSA in translation is a process of determining and maintaining coherence. Because coherence is a central issue to text comprehension, maintaining it in translation provides reader's model of representation of information as well as of their previous knowledge.

LSA can be used to identify locations in the text where topic shift occurs so that the text can be segmented into discrete topics (Landauer, Foltz and Kintsch, 1998). Discourse segmentation is based on the premise that the coherence should be lower in areas of discourse where the discourse topic changes.

Measuring the topic shift in machine translation is a big advantage which facilitates making more accurate text in TL applying LSA.

An LSA coherence analysis determines coherence entirely based on the derived semantic relatedness of one text unit to the next. Thus, it is making a coherence judgment based on the extent to which two text units are semantically related topic or have words that directly overlap. LSA does not perform any syntactic processing or parsing of the text. Within any unit of text, it does not take into account the order of the words. Despite not taking into account syntactic features, the analysis of the semantic features provide considerable strength in prediction. LSA captures Halliday and Hasan's (1976) notion of cohesion through lexical synonymy and hyponymy. In addition, it goes beyond this level in determining coherence based on semantic relatedness due to terms tending to occur in similar contexts (Landauer, Foltz and Kintsch, 1998), hence LSA makes machine capable of translation coherently into anther language. Although LSA lacks certain components of a cognition such as word order, syntax, or morphology, the representation it produces is highly similar to that of humans (Landauer & Dumais, 1997). By facilitating machine with syntax, etc along with LSA machine translation will show significant sign of improvement than without taking LSA into account.

Moreover LSA can be used to detect irony. From a discourse theoretic perspective irony means perceiving the distance between two points on a scale (Aynat, 2002)

Scale bottom **scale top**

Implicature from context **literal message**

In understanding an ironic utterance, one point is conveyed by the literal meaning of the utterance, and the other is a relevant implicature extracted from context. From a computational point of view, the quantitative gap between the literal and contextual meaning can be measured by LSA as the formal framework. LSA provides a metric that can be utilized to calculate the distance between the implicature and literal meaning. Machine translation will have more force to do the feat of recognition and render irony and metaphor which are calculated based on the distance between literal and non-literal meaning. The key idea in using LSA for translation is to look for dissimilarity and contrast which in LSA term means low similarity scores.

After all the use of LSA for machine translation should be tested thoroughly because varying the corpus on which LSA is trained may have a considerable effect on the result. Moreover, it is claimed that LSA represents words of similar meaning in similar way (Landauer et al, 1998) and is unable to detect synonyms from antonyms (Aynat, 2002), for this reason strategies should be taken into account to enable machine for distinguishing the two. It is also important to be aware that the relationships inferred by LSA are not logically defined, because they are relations only of similarity or of context sensitive similarity and so inferences extracted may give rise to fuzzy results that may be weak or strong.

Up to the present, it was argued that applying LSA engenders salient improvement in machine translation, but Hofmann (1990) introduces a novel technique called Probabilistic Latent Semantic

Analysis (PLSA) which has had strong impact on many applications ranging from information retrieval, information filtering and intelligent interfaces to speech recognition, natural language processing and machine translation. Both LSA and PLSA have the same idea which is to map high-dimensional vectors representing text documents to a lower dimensional representation called a latent semantic space (Kim, Chang, Zhang, 2002). PLSA is a technique for the analysis of two-mode and co-occurrence data. PLSA compared to LSA which is based on linear algebra and performs a Singular Value Decomposition of co-occurrence tables is based on a mixture decomposition derived from a latent class model. PLSA results in a more principled approach which has a solid foundation in statistics (Hofmann, 1990).

One of the fundamental problems is to learn the meaning and usage of words from some given corpus, possibly without further linguistic prior knowledge. The main challenge a machine has to address roots in the distinction between the lexical level of "what actually has been said or written" and semantic level of "what was intended" in a text or utterance. PLSA is more powerful in detecting polysemous words, i.e. a word which has multiple senses and multiple types of usage in different contexts (Hofmann, 1990), ergo PLSA can cope with translation more elegantly than latent semantic analysis (LSA).

The starting point for probabilistic latent analysis is a statistical model which has been called aspect model (Hofmann et al, 1999). The aspect model is a latent variable model for co-occurrence data which associates on unobserved class variable $z \in Z = \{z1,...., zk\}$ with each observation. A joint probability model over $D \times W$ is defined by the mixture

$$P(d,w) = p(d)\,p(w|d),\ p(w|d) = \sum_{Z \in Z} p(w|z)\,p(z|d).$$

like all statistical latent semantic variable models the aspect model introduces a conditional independence assumption, namely that d and w are independent conditioned on the state of the associated latent variable(Hofmann, 1990).

CONCLUSION

Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual usage of words by statistical computations applied to a large corpus of data. Its more powerful version probabilistic LSA is a new method, too. The use of these two techniques into translation will facilitate translation more automatically and accurately than without their application in MT. Applying these methods will produce texts in TL as they are comprehended by human being and used in different contexts.

ABOUT THE AUTHOR(S)

Mahmoud Mobaraki has academic studies in English literature as his B.A, theoretical linguistics as his M.A and applied linguistics as his PH.D. He is interested in machine and manual translation, dialectology and critical discourse analysis. He is the author of several articles published in reputed journals. Now, he is an assistant Professor in linguistics in Jahrom University, Iran.

Abolfazl Mosaffa Jahromi has academic studies in teaching English as his B.A, theoretical linguistics as his M.A and PH.D. He is interested in machine and

manual translation, syntax and morphology. He is the author of several articles published in reputed journals. Now, he is an assistant Professor in linguistics in Jahrom University, Iran.

REFERENCES

Aynat, R. (2002). *Using LSA to Detect Irony*. Tel Aviv University.

Bar-Hillel, Y. (1960). The present status of automatic translation of language. *Advances in computer,* 1, 91-163.

Britton, B. K. & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational psychology*, 83, 329-345.

Deerwester, S., Dumaise, S. T, Furnas, G. W, Landauer. T. K. and Harshman. R. (1990). Indexing by latent semantic analysis. *Journal of American Society for Information Science*, 41.391-407.

Dumais, S.T. (1994). Latent semantic indexing (lsi) and trec-2. *The Second Text Retrieval Conference (TREC2)*, D. Harman, editor, number 500-215 in National Institute of Standards and Technology Special Publication, pages 105–116.

Foltz, P. W, and Dumais, S. T. (1992). Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM*, 35(12):51-60.

Foltz, P. W; Kintsch, W and Landuer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes,* 25(2&3). 257-307.

Halliday, M. A. K. & Hasan, R. (1997). *Cohesion in English.* London: Longman.

Hofmann, T. (1990). *Probabilistic latent semantic analysis*. In fifteenth conference on uncertainty in artificial intelligence, (UAT99).

Hofmann, T., Puzicha, J. and Jordan, M.I. (1999). Unsupervised learning from dyadic data. *Advances in Neural Information Processing systems,* Volume 11. MIT press.

Hutchins, W. J. (1988). *Research methods and system designs machine translation: a ten-year review.* 1984-1994. in Clarke D. and Vella. A (eds) Machine translation: ten years on, procceding. 12-14 November 1994(Bedford: Granfield University Press, 1998), 4: 1-16.

Hutchins, W. J. (1986). *Machin translation: post, present, future. Chichester (UK)*: Ellis Horwood.

Kim, Y-Seop, Jeong-Ho Chang, and Byoung-Tak Zhang (2002). A comparative evaluation of data-driven models in translation selection of machine translation. *In 19th International Conference on Computational Linguistics*, pages 1–7, 2002.

Kim, W., and Sanjeev Khudanpur, (2004). Lexical triggers and latent semantic analysis for cross-lingual language model adaptation. *ACM Transactions on Asian Language Information Processing*, 3(2):94–112.

Jan, D. (2001). Special theme on machine translation. *International Journal of Translation,* 13. PP. 5-20.

Landuer, T.K. and Dumais, S. T. (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.

Landauer, T. K; Foltz, P. W and Laham D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2&3). 259-248.

Landauer, T. K; Foltz, P. W and Laham D. (2000). The intelligent essay assessor. *IEEE Transactions on Intelligent Systems*, 15(5):27–31.

McNamara, D. S. Kinntsch, E Songer, N. B. & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14, 1-43.

Newmark, P. (1981). *Approaches in translation***,** Oxford and New York: Pergamon.

Rehder, B., Michael L. Littman, Susan Dumais, and T.K. Landauer, (1998). Automatic 3-language cross-language information retrieval with latent semantic indexing. *The Sixth Text Retrieval Conference (TREC-6)*, Gaithersburg, MD, National Institute of Standards Technology.