Detection of Synonyms from Myanmar Text Using Latent Semantic Analysis

Myint Myint Win, Ni Lar Thein
University of Computer Studies, Yangon
myintmyintwin.ucsy@gmail.com

Abstract

Manually produced lexical resources are time-consuming and labour-intensive. Therefore, automatically identifying these resources is important to overcome knowledge bottleneck in natural language processing. This paper deals with the task of finding synonyms and near synonyms from raw data for the lexical substitution task. In this paper, an approach for automatically identifying synonyms is presented to support the process of semantic simplification for Myanmar text. Candidate synonyms for a target word will be generated using latent semantic analysis (LSA), which is a fully automatic statistical technique for extracting relations of expected contextual usage of words. This work can be helpful for thesaurus construction for Myanmar language.

1. Introduction

Synonym recognition is an important issue in a variety of fields dealing with language processing such as information retrieval and text summarization. Synonyms are different words with almost identical or similar meanings. For example, the words ကောန်, အမှန်, and စင်စစ် are synonyms [9]. In information retrieval, synonyms can be used to augment the query in order to improve the searched results of the retrieved articles [4, 7]. In summarization, synonyms are used for detection of redundant phrases in a text.

There are several domain independent lexical databases which provide short, general definitions of the words, and records the various semantic relations between the words for English and many other languages. Having lexical resources makes it easy to detect synonym words from the text. For Myanmar language, synonym detection is a challenging task due to its language-specific issues and lack of lexical resources.

Work in computational linguistics related to synonym detection has mainly focused on detecting semantically related words rather than exact synonyms, often using surrounding words to cluster words according to their similarity. The words are surrounded by context, which theoretically help in the process of synonym detection [8]. Context is the environment in which a word is used and it provides the information to figure out the meaning of a new or polysemous word.

There are a number of approaches for computing the semantic similarity between words. Reported approaches to solve synonym detection include latent semantic analysis (LSA), pointwise mutual information (PMI), matrix of proximity in documents combined with patterns of incompatibility, thesaurus-based methods, corpus-based similarity matrices, and a combination of various procedures.

The semantic similarity between words is based on the similarity between contexts. The vector space model is a popular technique to encode contexts and measure their similarity. For a target word which appears in a corpus, a context length, for example, the words in the same sentence or the words in a window of width *L*, is defined and all the words in the context of every occurrence of target word inside a bag. The several semantic similarity metrics can be defined between the bags corresponding to two words. There is much information lost in the vector space model because all the words are put together in the bag. Therefore, it can be extended with latent semantic analysis (LSA), a dimensionality reduction procedure in [17].

In this paper, a classical technique, latent semantic analysis (LSA), will be applied to detect automatically candidate synonyms from Myanmar text to support the lexical substitution task. It is a statistical method that extracts meaning of words using the information about the usage of the words in the context. LSA represents the text as a matrix and applies singular value decomposition (SVD) to that matrix. The result from singular value decomposition is a representation from which similarity measures between all pairs consisting of words or contexts can be calculated in the reduced dimensional space.

The rest of the paper is organized as follows. Section 2 reviews related work. The brief description of LSA and SVD is presented in section 3 and 4 respectively. The proposed system for detecting candidate synonyms from Myanmar text is presented in section 5, and finally end with conclusion in section 6.

2. Related Work

Much research has been carried on the search for semantically similar words in corpora in an automated manner in English and some other languages. However, a common problem is the size of the corpus available and for this reason they used to focus on restricted domains. One of the famous work proposed by [17] use document distribution to measure word similarity. They showed that LSA does not yield the best results when working with smaller document sets.

An unsupervised method which is based on the mutual information scores between a near synonym and the content words in the context filtering out the stop words is described in [1]. The pointwise mutual information (PMI) between two words compares the probability of observing the two words together i.e. their joint probability to the probabilities of observing each word independently.

The PMI-IR is a hybrid approach to deal with synonym detection [13]. It uses a combination of Pointwise Mutual Information (PMI) and Information Retrieval (IR) features. This work does not follow the attributional similarity paradigm but rather propose a heuristic to measure semantic distance.

The PMI-IR algorithm is refined and proposed a module combination to include new features such as LSA and thesaurus evidences [3]. A method which combines both approaches by employing global and local evidence of attributional similarity into a single measure is proposed in [15].

However, the most similar words produced using these methods are not always near synonyms, but may be words in other semantic relationships such as antonyms, hyponyms. Therefore, manual post processing of such automatically produced resources to filter out unwanted words may be necessary before they can be used [3]. They used Lin’s measure[2] of distributional similarity which is described by the ratio between the amount of information needed to state the commonality of the two words and the information needed to fully describe each word, to obtain most similar words to the target word and re-ranked the candidates using the overlap methods.

3. Latent Semantic Analysis

Latent Semantic Analysis (LSA) analyzes texts to find the underlying meaning or concepts of those texts. It keeps information about which words are used in the context such as a sentence or a passage, while preserving information of common words among sentences or passages, and represents the meaning of a word as a kind of average of the meaning of all passages in which it appears. It identifies the statistical association of words based on the assumption that there is some underlying latent semantic structure in the text.

The first step in LSA is to represent the text as a matrix in which each row stands for a unique word and each column stands for a text passage or other context. It then uses a statistical technique, called singular value decomposition (SVD) to estimate the latent structure. The results of this decomposition are descriptions of words and contexts based on the latent semantic structure derived from SVD. This structure is called the hidden concept space, which associates syntactically different but semantically similar words and contexts.

Let *D* be the text collection, the number of distinctive words in *D* be *m* and the number of text passages or other context in *D* be *n*. LSA starts with a $m \times n$ term-document matrix *M*. Each row of *M* represents a word and each column represents a context. Each entry of the matrix *M*, denoted by M_{ij}, is the number of times that word occurs in context *j*.

$$M = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

A row in this matrix will be a vector corresponding to a word, giving its relation to each context.

$$[a_{i1}, \ldots, a_{in}]$$

A column in this matrix will be a vector corresponding to a context, giving its relation to each term.

$$\begin{bmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{bmatrix}$$

LSA method then uses singular value decomposition (SVD) for finding out semantically similar words and passages.

4. Singular Value Decomposition

The singular value decomposition is a way of factoring matrices into a series of linear approximations that expose the underlying structure of the matrix. In SVD, a rectangular matrix is decomposed into the product of three other matrices: one component matrix describes the original row entities as vectors derived orthogonal factor values, another describes the original column entities in the same way, and the third is a diagonal matrix containing scaling values such that when the three components are matrix-multiplied, the original matrix is reconstructed.

The singular value decomposition of a $m \times n$ real or complex matrix *M* is a factorization of the form

$$M = U\Sigma V^T$$

where U is a $m \times r$ matrix and its columns, called right singular vectors, are eigenvectors associated with the *r* non-zero eigenvalues of MM^T. The columns of U are unit orthogonal vectors, i.e., $U^T U = I$ (identity matrix).

V is an $n \times r$ matrix and its columns, called right singular vectors, are eigenvectors associated with *r* non-zero eigenvalues of $M^T M$.

The columns of V are also unit orthogonal vectors, i.e., $V^T V = I$. $\sum$ is a $r \times r$ diagonal matrix, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_r), \sigma_i > 0$, $\sigma_1, \sigma_2, \dots,$ and σ_r, called singular values, are the non-negative square roots of *r* non-zero eigenvalues of MM^T. They are arranged in decreasing order, i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

5. Detecting Candidate Synonyms Using Latent Semantic Analysis

The purpose of the system is to find candidate synonyms from Myanmar text using Latent Semantic Analysis for the lexical substitution task. In Myanmar text, there are no explicit word boundary delimiters and therefore the text is segmented into words as the preprocessing step.

5.1. Word Segmentation

Word segmentation is the process of dividing written text into word boundaries. It is a fundamental and essential step in text processing. Word boundaries can be easily determined in English language by white space between words. In Myanmar language, the text is a string of characters written in sequence from left to right and words are not always delimited by spaces although sentences are clearly delimited by a sentence boundary marker “။” (ပုဒ်မ).

Therefore, the text is needed to be segmented into words as the preprocessing step in order to process the text computationally. Even though research have been carried for word segmentation, the resolution is not satisfying yet because of out of vocabulary words and over segmentation due to longest string matching [5].

To repair the errors caused by over segmentation, Conditional Random Fields (CRF) [6] is used. CRF++ is a simple, customizable, and open source implementation designed for a variety of natural language processing tasks and for segmenting or labeling sequential data. It can be downloaded at http://crfpp.sourceforge.net.

In CRF++, the feature template has to be specified in advance. The training and test file consist of multiple tokens. A token consists of fixed numbers of multiple columns. Each token must be represented by one line, and the columns are separated by the white space. A sequence of tokens corresponds to words. An empty line is put to identify the boundary between sentences.

Myanmar text are collected and applied to word segmentation algorithm [18]. The errors in the results of word segmentation algorithm are manually repaired and used as the training data for CRF++.

The following example shows a sentence and its segmented words.

မီးကိုပေါ့ပေါ့တန်တန်သဘောထားရန်မသင့်ပါ။

မီး_ကို_ပေါ့ပေါ့တန်တန်_သဘောထားရန်_မသင့်ပါ။

The sentence means that “Fire should not be considered carelessly”.

5.2. Constructing the Latent Space Representation

The first step to construct the latent space is to assemble the text of Myanmar language that is as similar as possible in size and content into meaningful passages such as sentences and paragraphs. Then, a matrix is created with words as rows and passages as columns.

To reduce the size of the problem, closed class words such as ၏, ကို, and ၌ are not considered in constructing the occurrence matrix. Each cell in the matrix contains the number of times that a given word is used in a given passage. An illustration of the system is shown with a small example which contains 21 words as rows and 7 sentences as columns.

$$M = \left[\begin{array}{ccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \begin{array}{c} \text{ဗီး} \\ \text{ပေါ့ပေါ့တန်တန်} \\ \text{သဘောထား} \\ \text{မထင်ပါ} \\ \text{လူကြီးများ} \\ \text{အစလေးအနက်} \\ \text{ပြောသော} \\ \text{စကား} \\ \text{ကပြက်ကချော်} \\ \text{မပြောသင့်ပါ} \\ \text{အစရော်ကြီးအသေး} \\ \text{ကိစ္စ} \\ \text{အရှုပ်ဖြီး} \\ \text{ဖြစ်မည်} \\ \text{နေမကောင်း} \\ \text{မကောင်းပါ} \\ \text{အလုပ်} \\ \text{လုပ်} \\ \text{လုပ်တတ်သော} \\ \text{ရှိသည်} \\ \text{မဖြစ်တော့ပါ} \end{array}$$

The matrix M is decomposed into three other matrices by using singular value decomposition according to the following equation

$$M_{m\times n} = U_{m\times k}\,\Sigma_{k\times k}V_{k\times n}$$

to represent the words and passages as vectors in a high dimensional semantic space. m and n are the number of words and passages respectively, and k is the number of dimensions for the latent semantic space.

$$U = \begin{bmatrix} 0.11 & -0.16 & -0.04 & 7.18 & -0.23 & -2.49 & -0.68 \\ 0.35 & -0.58 & 0.21 & 3.85 & -0.16 & -7.83 & 0.15 \\ 0.65 & 0.01 & 0.17 & 3.32 & 0.24 & 1.20 & -0.26 \\ 0.09 & -0.13 & 0.17 & 3.87 & 0.24 & -0.50 & 0.13 \\ 0.13 & 0.15 & -0.37 & -7.85 & 0.25 & 5.78 & 0.08 \\ 0.13 & 0.05 & -0.38 & -8.94 & 0.25 & 6.27 & 0.08 \\ 0.13 & 0.05 & -0.37 & -8.83 & 0.25 & 6.58 & 0.08 \\ 0.30 & -0.26 & -0.50 & -1.08 & -0.39 & -1.08 & -0.04 \\ 0.37 & 0.43 & -0.14 & -2.87 & -0.02 & -6.40 & 0.15 \\ 0.05 & -0.16 & -0.09 & -2.57 & -0.41 & -1.39 & 0.57 \\ 0.12 & 0.19 & 0.12 & 0.35 & -0.13 & -3.67 & 0.04 \\ 0.12 & 0.19 & 0.12 & 0.35 & -0.13 & -3.67 & 0.04 \\ 0.12 & 0.19 & 0.12 & 0.35 & -0.13 & -3.67 & 0.04 \\ 0.12 & 0.19 & 0.12 & 0.35 & -0.13 & -3.67 & 0.04 \\ 0.10 & 0.13 & 0.17 & 2.24 & 0.24 & 0.50 & 0.13 \\ 0.10 & 0.13 & 0.17 & 2.24 & 0.24 & 0.50 & 0.13 \\ 0.12 & 0.19 & 0.12 & -0.35 & -0.13 & -4.75 & 0.04 \\ 0.12 & 0.19 & 0.12 & -0.35 & -0.13 & -4.75 & 0.04 \\ 0.12 & 0.19 & 0.12 & -0.35 & -0.13 & -4.75 & 0.04 \\ 0.12 & 0.19 & 0.12 & -0.35 & -0.13 & -4.75 & 0.04 \\ 0.11 & -0.16 & -0.04 & 7.18 & -0.23 & -2.49 & -0.68 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 3.59 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.51 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.06 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.00 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.53 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.41 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.10 \end{bmatrix}$$

$$V - \begin{bmatrix} 0.33 & -0.33 & 0.35 & 7.34 & 0.37 & -0.71 & 0.15 \\ 0.48 & 0.13 & -0.77 & -1.81 & 0.38 & 8.81 & 0.09 \\ 0.20 & -0.40 & -0.18 & -3.50 & -0.62 & -1.94 & \\ 0.42 & 0.47 & 0.25 & 0.71 & -0.20 & -5.23 & 0.04 \\ 0.33 & -0.33 & 0.35 & 5.30 & 0.37 & 0.71 & 0.15 \\ 0.42 & 0.47 & 0.25 & -0.71 & -0.20 & -6.74 & 0.04 \\ 0.40 & -0.40 & -0.07 & -8.30 & -0.35 & -1.54 & -0.75 \end{bmatrix}$$

The matrix $U_{21\times 7}$ is a clustering of the words in the concept space. A word w is represented by the row in U corresponding to the row for w in M. The matrix $V_{7\times 7}$ is a clustering of the passages in the concept space. We can delete some insignificant dimensions in the concept space to optimally approximate matrix M.

5.3. Measure Similarity of Words

Once the semantic space has been created, the similarity of any two words or any two passages can be computed in the concept space. To calculate the similarity between two words which are represented by context vectors, cosine similarity measure is used. Each word is represented as a vector $\vec{x} = (x_1, x_2, \ldots, x_n)$ and similarities between such word representations will be calculated using the following equation.

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

Cosine is a commonly used similarity measure in which the two words are similar if their corresponding vectors are close to each other. The cosine ranges from 1, for perfectly correlated vectors, to 0 for totally uncorrelated vectors, and to -1 for perfectly inversely correlated vectors. For the sample text, ကပြက်ကချော် and ပေါ့ပေါ့တန်တန် are synonym words.

6. Conclusion

In this paper, we have presented an unsupervised, corpus-based statistical approach for detecting candidate synonyms from Myanmar text using Latent Semantic Analysis for the lexical substitution task. In general, LSA has good results in text data. However, different senses of words are treated as one and different word forms are treated as separate words. As an example, the word “ဘာသာ” have different senses in the sentences “ဘာသာနှင့်လူမျိုးကို ချစ်တတ်ပါ” which means that “Have affection for language and nationality” and “ငါ့ဘာသာ တစ်ယောက်တည်း သွားဝံ့သည်” which means that “I dare to go alone”. A large textual input to create a multi-dimensional semantic space is also needed.

References

[1] D. Inkpen, “A statistical model for near-synonym choice”, *ACM Transactions on Speech and Language Processing 4*, 2007

[2] D. Lin, “Automatic Retrieval and Clustering of Similar Words”, In Proceedings of COLING-ACL 98, Montreal, Canada, 1998

[3] D. McCarthy, B. Keller and R. Navigli, “Getting Synonym Candidates from Raw Data in the English Lexical Substitution Task”, 2009

[4] D. Sanchez and A. Moreno, “Automatic Discovery of Synonyms and Lexicalizations from the Web”, *In Proceedings of the 8th Catalan Conference on Artificial Intelligence*, 2005

[5] Hla Hla Htay and Kavi Narayana Murthy, “Myanmar Word Segmentation using Syllable Level Longest Matching”, *The 6th Workshop on Asian Language Resources*, 2008

[6] J. Lafferty, A. McCallum and F. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, 2001

[7] M. Baroni and S. Bisi, “Using Co-occurrence Statistics and the Web to Discover Synonyms in a Technical Language”, *In LREC*, 2004

[8] M. Ruiz-Casado, E. Alfonseca, and P. Castells, “Using Context-window Overlapping in Synonym Discovery and Ontology Extension”, 2005

[9] Myanmar Dictionary, Department of Myanmar the Language Comission, Ministry of Education, Union of Myanmar

[10] Myanmar Grammar Book, Department of the Myanmar Language Commission, Union of Myanmar

[11] O. Ferret, “Testing Semantic Similarity Measures for Extracting Synonyms from a Corpus”, 2010

[12] P. D. Turney, “Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL”, *In Proceedings of the 12th European Conference on Machine Learning (ECML-2001)*, 2001, pages 491-502

[13] R. Angheluta and M. Moens, “A Study About Synonym Replacement in News Corpus”, 2001

[14] R. Mihalcea, C. Corley and C. Strapparava, “Corpus-based and Knowledge-based Measures of Text Semantic Similarity”, *American Association for Artificial Intelligence*, 2006

[15] R. Moraliyski and Gael Dias, “One Sense per Discourse for Synonym Detection”, *In 5th International Conference Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, 2007

[16] T. K. Landauer, P. W. Foltz and D. Laham, “Introduction to Latent Semantic Analysis”, *Discourse Processes*, 1998, 259-284(6)

[17] T. Landauer and S. Dumais, “A Solution to Plato’s Problem: A latent semantic theory of the acquisition, induction, and representation of knowledge”, Psychological Review, 1997, 211-240

[18] Win Pa Pa and Ni Lar Thein, “Myanmar Word Segmentation Using Hybrid Approach”, 2008