An Ontological Approach to Analyzing Relational Similarities in the Yami Language

Meng-Chien Yang
Department of Computer and Communication Engineering
Providence University Taiwan

D. Victoria Rau
Institute of Linguistics, National Chung Cheng University, Taiwan

Abstract

Relational similarity is defined as the semantic similarity of a group of words (Turney 2006) measured by the corresponding relationship between pairs of words or phrases. Although many studies have proposed methods for finding relational similarities for major languages, such as English and Chinese (Nakov and Hearst 2008), there has not been much research on endangered languages. The biggest challenge for this line of research has been finding a proper corpus as comprehensive as Chinese Word Sketch (http://wordsketch.ling.sinica.edu.tw) to provide information on the structures of the word-pairs with synonyms and analogies. Recently, the Digital Archiving of Yami Language Documentation Project (http://yamiproject.cs.pu.edu.tw/yami) has accumulated a sizeable Yami corpus. This contribution has provided a foundation for ontological research.

This paper proposes a method for finding relational similarities in the Yami language. We begin by introducing the process of developing this corpus, followed by proposing a method to detect the relational similarity of pairs of Yami phrases by using the ontological structure of semantic representation and latent relational analysis (LRA) of the corpus. The proposed method can automatically derive the patterns of relational similarity from our corpus to find possible synonyms of the tested word pairs. After the ontologies of Yami phrases are built, an algorithm based on the LRA approach is used to find the relational similarity between the word pairs.

We will illustrate the detailed algorithms and our experimental results using the constructed dialogues about tourism on Orchid Island from our e-learning materials (http://yamiproject.cs.pu.edu.tw/elearn).

---

1 This research was supported in part by Taiwan E-learning and Digital Archives Programs (TELDAP), sponsored by the National Science Council of Taiwan under the NSC Grant: NSC 98-2631-H-126-001. The authors would like to express our gratitude for the editorial service of Dr. Gerald Rau.
I. Introduction

This paper aims to propose a feasible approach to automatically finding relational similarities in the Yami language, an endangered Austronesian language spoken by 4000 people on Orchid Island. Recently, more and more research projects on natural language processing have focused on similarity detection problems for various languages. There are two kinds of similarity for language processing, attributional similarity and relational similarity. Attributional similarity is regarded as highly similar relations between two words or two phrases. Usually, these two words/phrases are called synonyms. On the other hand, relational similarity refers to the relationship between two pairs of words. Each pair of words contains two words or two phrases. When the relation in one pair is identical to the relation in the other pair, these two pairs are said to have relational similarity, or they are analogous to each other. Analogue analysis is an important methodology for logical processing and language understanding. Furthermore, the algorithms for measuring relational similarity are useful and critical for web semantic applications because the web information retrieval process requires efficient algorithms to collect various semantic relations that connect words, phrases or languages in the huge world-wide-web.

Currently, many relational similarity measurement algorithms have been proposed to find word analogies from language corpora. However, according to the discussion of Turney (2006), the development of algorithms measuring relational similarity is not as comprehensive as the development of the algorithms for finding synonyms. Therefore, it would be beneficial to develop a new algorithm to measure relational similarity, as many new theories and findings on natural language processing can be collected from the process of algorithm development. In this paper, as we discuss these algorithms and the possible functions of relational similarity measurement, we will demonstrate that the development of a relational similarity measurement algorithm plays an important role in information extraction. However, most algorithms for measuring relational similarity have focused on English-related corpora. In this paper, the target language is Yami. Although the relational similarity measurement algorithms for English are our major reference studies, we are cognizant of the fact that Yami, an Austronesian language, is very different from English.
The goals of this study are two-fold. One is to explore the development of similarity measurement algorithms for an Austronesian language to establish a language web resource for it. The other is to explore the possibility of using relational similarity measurement algorithms for studies of other Taiwanese indigenous cultures.

The previous algorithms for measuring relational similarity have used one large training corpus for learning and another large corpus for testing. For example, many studies, such as Turney (2006), Turney (2009), and Kato et al. (2009), used huge word databases, e.g., the SAT testing data and WordNet (Fellbanum 1998). However, in our project, it is impossible to find a corpus containing such a huge number of words. To overcome the difficulties, we propose an algorithm using the ontological structures to extract semantic vectors to find the relational similarity measurement.

We will present the experimental data and possible applications. As with previous studies of algorithms, this study started with a set of Yami queries to find possible sets that contain the words with the predefined relational measurements. We will demonstrate that our preliminary results indicate the proposed algorithm is useful in collecting pairs of relational similarity in the Yami language.

The paper is organized as follows. In Section II, the algorithms for relational similarity are described. Section III shows our proposed methods for finding relational similarity in the Yami language in the process of documenting the Yami language. Section IV illustrates our algorithms for finding relational similarity. The results and discussion are in Section V. The conclusions and future studies are in Section VI.

II. Related Works

Relational similarity measurement for language processing is a fairly new research topic. Previous studies can be classified into two categories.

The first approach is based on the Vector Space Model (VSM). The VSM-style approach used a large corpus to train a dataset that contains a set of vectors representing the frequencies, relations and orders of the words in the corpus. The training process includes two major steps. In the first
step, a large set of queries are used to find the matching words in the corpus. In the second step, the results of query matching are transferred to the vectors of the dataset. When a new word query is put into the system, the dataset is used to calculate the relational similarity measure values. The measuring algorithms are developed using these vectors and related information. A structure called Latent Relational Analysis (LRA), proposed by Turney (2006) and Turney (2009), is used to automatically compare patterns of words to find the contextually synonymous information. In the LRA, a word pair is represented by a set of vectors of lexical patterns and frequencies. When there is a new query containing a set of words, a new word pair is created, compared with the dataset, and used to find possible candidates with the n-gram vector which is similar to the new word pair.

The VSM and LRA have been reported to have a precision rate of 56.4% in comparison with the SAT analogy word lists. However, the training required for LRA is very time-consuming. In addition, the dataset needs to be re-trained to a new dataset when there is a query from a new corpus. Therefore, there are great opportunities for improvement in future studies.

The second approach is based on the structure of the tested corpus. Veale (2004) developed an algorithm that used the structure and taxonomy in WordNet (Fellbanum 1998) to find possible analogies. The path of the semantic relations of words in WordNet is self-explanatory. However, WordNet is restricted to a certain domain of knowledge, and thus cannot be used to find the relational similarity of some specific domains, such as country names or city names.

Recently, another approach has been developed by mixing the structure of the tested corpus and the vector space. Bollegala et al. (2009) proposed to use lexical patterns to represent the structure of the words, use the clustering method to find the relational similarity between these word pairs, and finally use a metric similar to the one in VSM to find the relational similarity measurement values. This mixed approach is more suitable for developing the relational similarity measurement algorithm for our project. The details of the proposed algorithm will be described in Section IV.
III. Methodology and Yami Documentation

It is critical to develop a web language resource for endangered languages for the following reasons. First, many endangered languages are likely to be extinct in a decade or so. The scarce web language resources for the endangered languages make them totally invisible in the web community. Furthermore, this situation can even worsen the endangerment of those languages because the Internet becomes virtually inaccessible to people using the endangered languages.

Most of the endangered languages are neither well studied nor well documented. Most endangered language documentation projects have not considered the establishment of a web language resource as an aim of their projects (Birds, Simons and Huang 2001; Huang, Vogel and Waibel 2003). The problems for creating web resources for endangered languages include at least the following:

- The data and corpora are not comprehensive enough to form a complete language resource.
- The morphosyntax and semantics of the endangered languages are still under investigation.

Therefore, it is virtually impossible to establish a web language resource for endangered languages. Furthermore, compiling an endangered language resource requires knowledge and support from linguists, community members who are still using the endangered languages, and computer scientists.

Our Yami language research team presents a unique case, as we have documented and archived a reasonable size Yami corpus since 2005 as a result of the ELPD grant awarded to the two authors (http://yamiproject.cs.pu.edu.tw/yami/eng/). Recently, we have been exploring a proposed methodology to deploy our Yami resource as a web language resource with semantic relations, similar to WordNet (Fellbanum 1998), and discovered an approach and a platform with sound results.

We have developed a conceptual framework for creating the web resource for endangered languages, as shown in Figure 1. There are two processes in the framework. The first is to find semantic similarities between the endangered language and a complete web resource of a major
language. WordNet is selected as the complete web resource in the following illustrated figure. This process computes and extracts sentential and semantic similarities between the major language and the endangered language.

The second process involves transforming the extracted web resources into a new endangered language web resource, as shown in Figure 1. The extracted resource may be open resources useful for the people using the endangered languages or the most common terms used in social networking in the Internet environment. The Yami resource can either provide useful information for Yami speakers or serve as an agent to identify technological terms that need to be translated from the major languages (e.g., English and Chinese) into the endangered language for educational purposes.

![Figure 1. A framework for creating an endangered language web resource for Yami.](image)

Establishing this framework for developing Yami language web resources requires many different tasks. The process of finding relational similarities in Yami is certainly an important part of the whole plan.

**IV. Relational Similarity Measurement Algorithm**

In this section, we describe the proposed relational similarity measurement algorithm. The proposed algorithm follows Turney's (2006) LRA approach. The function of the relational similarity measurement is described as follows: given a pair of input Yami words (A, B), the similarity measurement algorithm will find the relationship of sets of word-pairs to (A, B). As with other algorithms, there is a reference database containing possible pairs of a wordlist with
relational similarities. The proposed algorithm is described as a two-phase process: (1) learning phase and (2) similarity analysis phase.

The learning phase is used to derive relational similarities from a set of words from the Yami corpus. The Yami corpus used in our study is the Yami lexical database created by the Yami Documentation project team (Dong, Rau and Chang 2010; Rau et al. 2009; Rau and Yang 2009; Rau, Yang and Dong 2007; Rau and Dong 2006). The patterns of analogy are designed as an ontological structure in that the patterns are transferred to the ontological links defined by the OWL languages. The patterns selected as analogy patterns for Yami are the following:

- Patterns used for representing the analogies in WordNet,
- Patterns with cultural metaphors. In this study, we only use the cultural metaphors found in our previous study (Tai, Rau and Yang 2008),
- Patterns acquired from the contextual and relational similarity analysis.

The contextual and relational similarity analysis follows the algorithm proposed by Pei et al. (2006) and Bollegala et al. (2009) to process twelve selected Yami texts on our web site (http://yamiproject.cs.pu.edu.tw/yami/) to find relational similarities of the Yami words.

In the learning phase, we use the following steps to build the Yami relational similarity measurement database.

- Put all the words in the Yami lexical database into the measurement list.
- For each word in the measurement list, the patterns of the word compared with other words are established by filtering their semantic relations in the WordNet synonym list. The patterns are described as a set of vectors.
- For each word in the measurement list, the words having metaphoric links with other words are selected. The word pairs are grouped and transformed into vector patterns.
- For each word appearing in the twelve Yami texts, the contextual and relational similarity analysis is conducted to find possible candidate words. The relations of these words with the measured word are transformed into vector patterns.
- All the vector patterns calculated in Steps 2, 3, and 4 are grouped as matrixes representing the similarity measures of this word.
- The matrixes for all words are built and put in a relational database.
Each vector represents the similarity measurement of a word with another word. In addition, we put the ontological information of these word pairs into the vector. In the similarity analysis phrase, these matrixes and vectors are used to find the relational similarities for an input Yami word pair \((X, Y)\).

The similarity analysis phase is designed as a system for processing the possible input from the web page. As the similarity analysis is presumed to find the input Yami words, the algorithm used in this phase should consider the execution time necessary.

Suppose that there is a pair of input Yami words \((X, Y)\): the similarity analyzing phase will retrieve a set of possible candidate list:

\[
S' = \{(A', B'), \text{where } R(A', B') \text{ has the same relationship as } (X, Y)\}.
\]

The steps of retrieving the candidate list are described as follows:

- Search for synonyms of each word and find the representational matrix of each synonym. This produces two synonym sets \(S(X'), S(Y')\) as the synonyms of \(X\) and \(Y\), respectively.
- For each synonym in \(S(X')\) and each synonym in \(S(Y')\), the similarity measurement is calculated using the matrix and its mapping to SVD. An entropy value is calculated as the similarity measurement. Then, a sorted list with the highest entropy is obtained as the result.
- For each selected word, the vector of the word is used to project possible relationships. We will use the vector to find more candidates that are missed by the entropy search in Step 2.
- A possible relational similarity list is found. Each pair \((X', Y')\) in the list is given a relational similarity value \(V\). The \(V\) represents the degree of relational similarity with the input word pair \((X, Y)\).

V. Results

This section presents results of our preliminary experiments with the input words from the Yami e-learning materials. The learning phrase is still ongoing in our system. For this test, we built a smaller word list from Book One (Beginning Level) of the Yami e-learning materials at
The word list was used to build the matrix using MATLAB.

After the matrix was built, a set of three Yami word pairs was used for testing. These word pairs were (rayon, vehan), (nohon, apia), and (kaka, wari). These words were selected from a pool of words related to the meaning: flying fish, good, and relatives. The selection proceeded randomly. The translations of these word pairs are (flying fish season, month), (good/okay, good), and (older sister/brother, younger sister/brother), respectively. The results of the experiment are shown in Table 1. In this table, the possible matching pairs were found in the WordNet corpus as English word pairs.

The results show that the proposed algorithm can find the structure and relations of the analogies. The matching rates are expected to improve if more word lists are put to the test, as will be tested in a future study. The results also revealed that the selection of the tested word pairs requires their lexical meanings as supplemental information. On this test, the word pair (nohon, apia) belongs to different word classes. The homophony of apia was also not controlled. Future studies on measurement of Yami word pairs will redress these problems.

Table 1. Results of the relational similarity algorithm.

<table>
<thead>
<tr>
<th>Input Word Pair</th>
<th>Output Word Pair (English)</th>
<th>Precision (matching rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(rayon, vehan)</td>
<td>(season, month)</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td>(flying fish, month)</td>
<td>30%</td>
</tr>
<tr>
<td>(nohon, apia¹)</td>
<td>(yes, egg)</td>
<td>20%</td>
</tr>
<tr>
<td>(okay, okay)</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>(right, right)</td>
<td>20%</td>
<td></td>
</tr>
</tbody>
</table>

² Apia has two homophones. Apia¹ means “good”; apia² means “egg.”
VI. Conclusion

The study is a new attempt to find possible approaches to searching the semantic structure of the Yami language. In this paper, a relational similarity measurement algorithm was proposed and tested in a preliminary experiment.

In future studies, we will carry out further experiments to test the validity of the Yami relational similarity algorithm. The establishment of a relational similarity algorithm will pave the way for developing related methodologies and language processing algorithms leading to building other Austronesian web resources.

References

Bird, S., Simons, G., and Huang, C.-R.  

Bollegala, D., Matsuo, Y., and Ishizuka, M.  

Byrne, A.  

Dong, M.-N., Rau, D. V., and Chang, H.-H.  

Fellbanum, C.  

Huang, F., Vogel, S. and Waibel, A.  

Kato, M. P., Ohshima, H., Oyama, S., and Tanaka, K.  
2009 Query by analogical example: Relational search using Web search engine indices, Proceedings of CIKM'09, pp. 27-36.

Pei, J., Han, J. Mortazavi-Asi, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., and Hsu, M-C
Rau, D. V., Yang, M.-C., Chang, H.-H., and Dong, M.-N.

Rau, D. V. and Yang, M.-C.

Rau, D. V., Yang, M.-C., and Dong, M.-N.

Rau, D. V. and Dong, M.-N.

Razavian, N. S. and Vogel, S.
2009  The Web as a platform to build machine translation resources, Proceedings of IWIC'09, pp. 41-45

Su, L. I.

Tai, Y.-S., Rau, D. V., and Yang, M.-C.

Turney, P. D.

Veale, T.