Language, Information and Computation(PACLIC 11), 1996, 373-382

Estimating Point-of-View-based Similarity using POV Reinforcement and Similarity Propagation

Kenji Nagamatsu Hidehiko Tanaka
The University of Tokyo
{naga,tanaka}@mtl.t.u-tokyo.ac.jp

Abstract

This paper proposes a similarity measure which takes account of point-of-views (abbreviated to POV, hereafter) in the calculation of similarity values. So far many researches on similarity measures have been performed but none takes account of POVs. The similarity measure proposed in this paper is based on co-occurrence probabilities of words and this makes it possible to obtain preferable precision even if POVs are not given. This method consists of two parts of processes, POV reinforcement and similarity propagation. First, the POV reinforcement process, which affects the similarity between words, modifies the weights of links according to the relatedness between the link and the POV word. Second, the similarity propagation process propagates the weights of links and defines a similarity value for word pairs which do not actually co-occur in the corpus. Using those two processes this method becomes capable both to take POVs into consideration and to cope with the sparseness of corpora to some degree. This paper, however, focuses on the POV reinforcement and evaluates the effectiveness of the method..

1 Introduction

Rapid growth of computer networks has increased the number of machine-readable texts and also made it possible for us to use various search engines to get desired documents. They are, however, keyword-based and strict. Even if some documents are related to a user's interest, he or she cannot obtain them as long as they do not contain the keywords given in advance. Otherwise he or she will be handed too many documents which contain just the keywords but are not necessarily related to his or her interest.

To solve these problems, it is necessary to take account of similarity between words or concepts and to make use of the measure in search processing. However, because there are many similar words in a text, employing a similarity measure alone only will expand the range of relatedness and produce more and more results.

When considering meanings of words, we, human beings, do not consider the whole meanings at a time, rather some interesting aspects of their concepts just the same as we look at a landscape from some point-of-view. Hence in similarity of words it is required to take account of their POVs. This makes it possible both to expand the range of matching in some situations and to restrict the range in others. Expectation is that employing valid POVs has both the expansion and the restriction be suitable and search processing produces more appropriate results.

This paper proposes a similarity measure between words which measure takes account of the effect of POVs. So far many researches on concept (or word) similarity have been performed but none handles POVs in their similarity measures. The proposed method utilizes co-occurrence probability-based similarity as a basis

and extends this fundamental measure by weighting the values according to the relevance between input words and POV words. This fundamental measure and its evaluation with some traditional similarity measures are described in 2. The main part of the method, which handles the effect by POVs, consists of two processes, *POV reinforcement* and *similarity propagation.* The explanation for these processes and some related issues are presented in 3 and 3.2. Finally 4 shows the result of some experiments, which indicates the effectiveness of this method, and 5 discusses the problems of the method as well as its advantages.

2 Fundamental Similarity Measures

This section gives an overview of similarity measures (2.1) and evaluates the ability of some fundamental measures with a large amount of word pairs (2.2 and 2.3).

2.1 Classification of Similarity Measures

Similarity of words or concepts is a fundamental measure in natural language processing because it can be used in various processing. For example, in disambiguation of word senses it can help to detect appropriate word senses by selecting most similar word senses to the senses of context words. In sentence production similarity measures can also help to keep coherence of word sequences. The fact that most researches on similarity measures have been performed with relation to the word sense disambiguation indicates the significance of similarity measures.

The similarity measures researched so far are classified as follows.

1. Similarity based on the structure of thesauruses or taxonomies (Agirre 1995)(Resnik 1995)
   Because thesauruses or taxonomies contain even infrequently used words (or concepts), the similarity measures of this type can define similarity values to most word pairs. The range of the word pairs they can handle is, thus, broad. In contrast, these measures are also considered as class-based and the degree of similarity is rather loose (they tend to judge the words or concepts in a same class as similar).
2. Similarity based on the statistical information extracted from corpora (Dagan 1994)(Iwayama 1994)(Yang 1994)(Karov 1996)
   The range of the word pairs they can handle depends on the size of corpora used to extract the statistical information. In most cases, however, the problem of data sparseness arises. The main concern in these measures is how to estimate the values of unseen word pairs (Dagan 1994).
3. Similarity based on network structures (Kozima 1993)(Niwa 1994)
   Similarity values are defined on links in the network and a total value in a path, maybe processed somewhat, is interpreted as a similarity value.
4. Feature-based similarity
   In these measures each word (or concept) has a set of features which are semantically related to the word. Those features may be actual semantic features or co-occurrence words. The number of shared features is interpreted as a similarity value.

2.2 Selectivity of Similarity Measures

Before considering the similarity measure based on POV, it is necessary to clarify the ability of some fundamental similarity measures described in the previous section. With the result this evaluation the most promising measure is adopted as the base of the proposed method.

In many researches the evaluation of similarity measures depends on human beings' judgment. In this case the measures are estimated by the score subjects judged for output values of the measures or by the correlation between two judgments of subjects and similarity measures. Scoring similarity of word pairs by hand, however, costs a lot. In contrast to this approach this paper adopt another type of evaluation employing coverage and selectivity of similarity measures.

When some threshold of a similarity measure is determined, the measure can judge each pair of words as similar or as not similar. At this point the coverage of a word pair set by the similarity measure is defined as the proportion of the number of word pairs judged as similar to the size of the set (total number of word pairs in the set).

Employing this coverage ratio, selectivity of a similarity measure is described as follows. First, two groups of word pairs are prepared. One group, *synonym set*, contains pairs of synonyms which are similar in human beings' judgment. The other group, *non-synonym set*, contains pairs of non-synonyms which are not similar to each other. In practice, however, non-synonym set is approximated with word pairs randomly selected from a dictionary. When some threshold of a similarity measure is determined, two coverage ratios for those two sets can be computed and the relationship between the two coverage ratios is plotted with the threshold being a parameter (see Figure 1 and 2 as examples). This plotted relationship is defined as the selectivity of the similarity measure. In the graphs, the lower a data sequence is located, the higher the selectivity of the similarity measure becomes.

2.3 Evaluation of Fundamental Similarity Measures

In this section some fundamental similarity measures are evaluated employing the selectivity. Evaluated measures are three, **depth**, **link#**, and **cooccur**. These are not the latest measures but are commonly used and offer bases of more advanced measures.

depth represents the similarity measure which uses the depth of the most specific common ancestors(MSCA). Given two concepts, MSCA are the concepts which subsume both the concepts and are located at the deepest position in a taxonomic structure. Formally,

$$Sim_{\mathbf{depth}}(w_1, w_2) = \max_{\forall c_1 \in C(w_1), \forall c_2 \in C(w_2)} \frac{d(\mathrm{MSCA}(c_1, c_2))}{(d(c_1) + d(c_2))/2} \quad (1)$$

, where $C(w)$ denotes the concept set of a word w and $d(c)$ denotes the depth of a concept c in a taxonomy.

link# represents the traditional edge counting method, which define the similarity value of a word pair (w_1, w_2) by the length of the shortest path from one

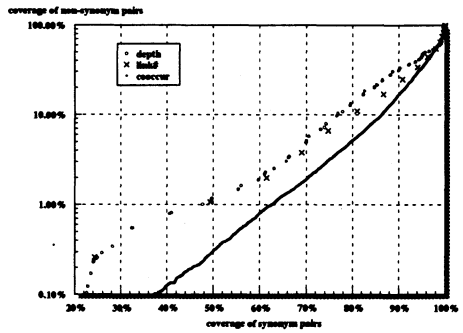

Figure 1: Selectivity of fundamental similarity measures

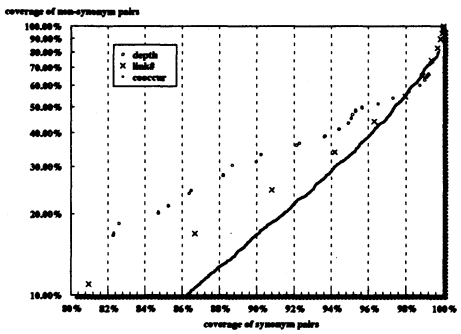

Figure 2: Magnified version of Figure 1

of the concepts of w_1 to one of the concepts of w_2. Formally,

$$Sim_{\texttt{link\#}}(w_1, w_2) = \max_{\forall c_1 \in C(w_1), \forall c_2 \in C(w_2)} \frac{1}{l(c_1, c_2) + 1} \tag{2}$$

, where $l(c_1, c_2)$ denotes the shortest path length between concepts c_1 and c_2 in a taxonomy.

cooccur represents the similarity measure which uses co-occurrence probability between words. Formally,

$$Sim_{\texttt{cooccur}}(w_1, w_2) = \sum_{\forall w \in Co(w_1) \cap Co(w_2)} \frac{Pr(w|w_1) + Pr(w|w_2)}{2} \tag{3}$$

, where $Co(w)$ denotes the co-occurring words with a word w and $Pr(w'|w)$ denotes the co-occurrence probability of w' conditioned by w. This measure has a name "cooccur" but is a hybrid type of statistics-based and feature-based similarity.

Figure 1 and 2 show the result of the evaluation employing the selectivity.

In this evaluation the synonym set contains 10,297 synonym pairs which was extracted from the IPAL dictionaries (IPA 1993) , which have a "synonym words" field in the word records.

Taxonomy-based similarity measures (**depth** and **link#**) use as a taxonomy the EDR concept dictionary (EDR 1995). The non-synonym set used for these measures are approximated with word pairs randomly selected from the EDR word dictionary (EDR 1995) , which contains the word entries corresponding to the concepts in the EDR concept dictionary.

For the co-occurrence-based similarity measure (**cooccur**), co-occurrence data were extracted from the corpus CD-Mainichi shimbun (newspaper) DB '94, which contains all the articles from this newspaper in 1994. This co-occurrence data contains the co-occurring words and their frequencies for each content word (nouns, verb, adjectives, and adverbs) in the corpus. The number of the sentences used for the extraction is 1,019,997(74,793 articles).

Figure 1 and 2 indicates clearly that in taxonomy-based similarity measures (**depth** and **link#**) the edge counting method is superior to the depth measure.

Moreover, the corpus-based similarity, which uses co-occurrence probability extracted from a corpus, is superior to the taxonomy-based measures.

(Resnik 1995) have extended the depth-based similarity measure by employing information content of concept classes, which was calculated from word frequencies in a corpus, and concluded that the method was superior to the edge counting method. On the other hand, the edge counting method has been extended to a network-based similarity model described earlier. In this way the combination with statistical information extracted from corpora produces preferable results. The POV-based similarity method presented in the next section also adopt the co-occurrence probability-based similarity as a basis.

3 Similarity of Words based on POV

This section describes the similarity measure proposed in this paper, which can take account of the effect of point-of-views. This similarity measure consists of two phases, *POV reinforcement* and *similarity propagation.* However, this paper focuses on the POV reinforcement and omits the explanation of the similarity propagation process. Before describing the POV reinforcement process (3.2) a similarity network with POV, on which the similarity measure is defined, and the method of calculating similarity values are explained.

3.1 Similarity Network with POV

As described in 1, human beings' judgment of similarity takes POVs into consideration. Two different words may not be similar in general, rather they are similar under some aspects or POVs. Thus we consider similarity of words as a triplet $Sim(w_1, w_2; w_p)$, where w_1 and w_2 are called node words (similarity values are defined over them) and w_p is called a POV word.

From this point of view the co-occurrence data used in 2.3 can be also used as the triplets because the co-occurring words of a node word are thought as POVs of the node. But if the co-occurring words are used as POV words directly, the sparseness problem arises because the co-occurring words don't necessarily contain the POV word given to the calculation. The POV reinforcement process, therefore, employs another type of co-occurrence data. The details will be described later in 3.2.

Even if the co-occurrence data cannot be used for the handling of POVs, these data can be used for the calculation of basic similarity values. As described in 2.3 the measure utilizing these data has higher selectivity than the other taxonomy-based measures. Therefore, as a fundamental structure the similarity measure defined by the equation (3) is adopted. $Sim(w_1, w_2; w_p)$ is, thus, formulated as follows.

$$Sim(w_1, w_2; w_p) = \sum_{\forall w \in Co(w_1) \cap Co(w_2)} \frac{Pr(w|w_1; w_p) + Pr(w|w_2; w_p)}{2} \qquad (4)$$

, where $Pr(w|w'; w_p)$ denotes the co-occurrence probability of w conditioned by w_1 which probability is reinforced by a POV word w_p. This reinforcement is described in the next section.

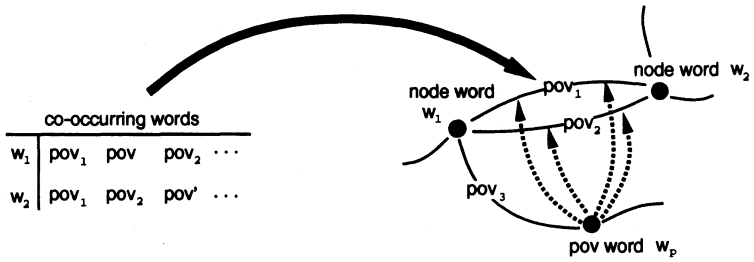

Figure 3: Similarity network with POV and POV reinforcement

The similarity network with POV is constructed as follows. First, the nodes of the network are the words which appears in the co-occurrence data described in 2.3. Second, every pair of nodes is connected by links which are correspond to the shared co-occurring words ($Co(w_1) \cap Co(w_2)$) respectively and each link is given a pair of the co-occurrence probability, one for w_1 and the other for w_2(see Figure 3).

3.2 POV Reinforcement

POV reinforcement is the most important process in this similarity measure and responsible for varying the values of links according to the relatedness to a POV word.

As described above, the co-occurrence data used in equation (4) cannot be employed to weight values of links according to POVs. It is because normal co-occurrence data are collected ignoring the relationship between co-occurring words. As a result a pair of words shares various POVs in the co-occurrence data. Therefore, another type of co-occurrence data, called POV co-occurrence data, is required.

To extract POV co-occurrence data from a corpus, we make two assumptions. 1) Two words are similar when they occur as the same case role of the same word (verb, etc.). 2) The POV of this similarity is the verb, etc. itself. For example, in two sentences 1) "Tom walks." and 2) "A dog walks.", both 'Tom' and 'dog' have occurred as *agent* of the verb 'walk'. 'Tom' and 'dog' are, thus, considered to be similar under the POV word 'walk'.

Following the assumptions, POV co-occurrence data in the form $co(w_p, w_i, r_k)$ are extracted from a tagged corpus. This gives co-occurrence frequency that word w_i occurs as the case role r_k of the word w_p. Employing these data the POV reinforcement is formulated as follows.

$$Pr(w'|w; w_p) = \frac{\alpha^{mic(w_p,w')} f(w'|w)}{(\alpha^{mic(w_p,w')} - 1) f(w'|w) + \sum_{\forall x \in Co(w)} f(x|w)} \quad (5)$$

, where $f(w'|w)$ denotes the normal co-occurrence frequency of w' conditioned by w and $mic(w_p, w')$ is the mutual information content which is calculated with

POV co-occurrence data. α is a constant parameter which controls how the relatedness between two POVs w_p and w' affect the probability of the link. This mutual information content $mic(w_p, w')$ is approximated as follows with POV co-occurrence data $co(w_p, w_i, r_k)$.

$$MIC(w, w') = \log \frac{Pr(w, w')}{Pr(w)Pr(w')}$$
$$\approx mic(w, w') = \log \frac{\sum_k co(w, w', r_k)}{\sum_{i,j} co(w, w_i, r_j) \sum_{i,j} co(w', w_i, r_j)} \quad (6)$$

Mutual information content(MIC) indicates the degree of co-occurrence. If $MIC(w_1, w_2) \gg 0$, the relationship between w_1 and w_2 is quite meaningful. If $MIC(w_1, w_2) \approx 0$, w_1 has nothing to do with w_2. And if $MIC(w_1, w_2) \ll 0$, w_1 and w_2 occur exclusively. This behavior of MIC is useful for weighting links according to the relatedness between POVs and links.

When no POV word is given, the equations (4) and (5) become the same as (3). This guarantees that this similarity measure has at least the same ability shown in the Figure 1 and 2 (cooccur).

4 Experiments

For these experiments POV co-occurrence data were extracted from the EDR corpus (EDR 1995). This corpus contains 207,802 sentences and all the sentences are already parsed into semantic frames. From these frames 1,254,851 POV co-occurrence data *co* are obtained. The normal co-occurrence data are the same as the data described in 2.3, which were extracted from CD-Mainichi shimbun (newspaper) DB '94.

4.1 Selectivity of the Measure with the POV reinforcement

This experiment evaluates the effectiveness of the POV reinforcement process. Because the case where POV words are given explicitly is difficult to control the conditions of the experiment, this experiment evaluated the case explicit POV words are not given. As described earlier, even if no explicit POV words are specified, the words in a input pair are used as implicit POV words.

Figure 4 and 5 shows the result obtained in the same way as 2.3. In the figures the multiple versions of the POV-based similarity measure are plotted at $\alpha = 1.2$, $\alpha = 1.5$ and $\alpha = 2.0$. α is the parameter of equation (5).

Table 1 contains coverage of non-synonym pairs for some typical coverage of synonym pairs.

4.2 Comparison with Human Judgment

The evaluation by the selectivity of the similarity measures is comprehensive but only suggests an overall tendency. Therefore, another experiment has been performed. This experiment compares the similarity values computed by the similarity measures with the scores given by subjects.

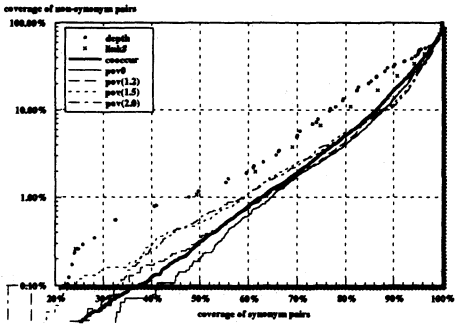

Figure 4: Selectivity of the proposed similarity measure

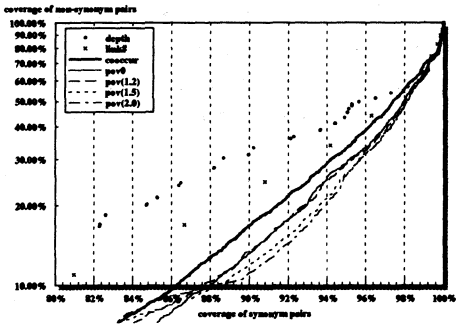

Figure 5: Magnified version of Figure 4

coverage of synonym pairs	POV $\alpha = 1.2$	POV $\alpha = 1.5$	POV $\alpha = 2.0$	cooccur	link#
80%	4.4%	5.6%	5.1%	5.2%	10.9%
90%	13.4%	11.9%	11.1%	16.8%	24.0%
95%	28.8%	26.6%	26.0%	33.0%	33.9%

Table 1: Coverage of non-synonym pairs for typical coverage of synonym pairs

The number of subjects was 14 and they were all members of our laboratory. They were asked to rate the similarity of each pair of words from 1 (not similar) to 5 (perfect synonymy). The number of pairs of words was 50, which were randomly selected from synonyms in the IPAL dictionaries.

Table 2 shows the result and contains the correlation factor between the values of the similarity measures and the scores given by the subjects.

5 Discussion

5.1 Effectiveness of the POV reinforcement

The result of this experiment (see Figure 5) indicates that by employing the POV reinforcement the selectivity of the measure becomes higher than the original cooccur measure (note again that the lower a sequence is located, the higher the selectivity of the measure becomes). This raise originates in the effect of POVs

$\alpha = 1.2$	$\alpha = 2.0$	cooccur	link#	depth
0.2051	0.1909	0.1987	0.0822	0.1277

Table 2: Correlation between the similarity measures and the judgment by subjects

alone. Although the measure with the POV reinforcement becomes inferior to the original one in the area where coverage of synonym pairs is small ($\approx 50\%$), this is not a problem because similarity measures are used normally at high synonym coverage ($\approx 80\% \sim 90\%$).

The effect of the parameter α is quite interesting. In proportion as α increases the selectivity also rises in the neighborhood of $90\% \sim 92\%$. However, in the area below 80% the selectivity becomes declined conversely. This is also observed more clearly from Table 1. It is considered that there is a optimum value of α, however it is not yet found.

5.2 Comparison with Human Judgment

Table 2 indicates that judgment by the co-occurrence-based similarity measures resembles that of human beings more than the taxonomy-based similarity measures. All the factors are, however, very small. (Resnik 1995) have presented the result of a similar experiment to this. There the correlation factor between the human judgment and the values of a edge-counting method is 0.6645. In comparison with the result in Table 2, because word pairs used in this experiment were selected from synonym pairs, it is considered that the difference among the similarity values became small.

Moreover, no effect of the POV reinforcement on these correlation factors is recognized. Considering its effect on the selectivity this is considered to be caused by the process of this experiment.

In either case it is required to perform another experiment thoroughly.

6 Conclusion

This paper has presented the similarity measure which takes account of point-of-views, focusing on the POV reinforcement process. Although this method consists of two phases, the POV reinforcement and the similarity propagation, the POV reinforcement process is the main part, which weights the co-occurrence probabilities of links according to POV words. The result of the evaluation suggests that the POV reinforcement have a good effect on the similarity measure. On the other, however, the experiment of comparison with human judgment did not produce satisfactory result.

As a future work, the thorough comparison with human judgment and the evaluation of the similarity propagation process are required. In addition, it is necessary to evaluate the behavior of the results when this similarity measure is used in practical processing, for example word sense disambiguation.

References

Eneko Agirre and German Rigau. 1995. A Proposal for Word Sense Disambiguation using Conceptual Distance. In *Proceedings of 1st International Conference on Recent Advances in Natural Language Processing.*

Ido Dagan and Fernando Pereira. 1994. Similarity-Based Estimation of Word Cooccurrence Probabilities. In *Proceedings of ACL-94*.

Japan Electronic Dictionary Research Institute Ltd. 1995. EDR Electronic Dictionary Technical Guide.

Japan Information-technology Promotion Agency. IPAL Japanese Dictionary for Computer. Technical report.

Yael Karov and Shimon Edelman. 1996. Learning similarity-based word sense disambiguation. In *Proceedings of the Fourth Workshop on Very Large Corpora.*

Hideki Kozima and Teiji Furugori. 1993. Similarity between Words Computed by Spreading Activation on an English Dictionary. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL-93)*, pp. 232–239. ACL.

Yoshiki Niwa and Yoshihiko Nitta. 1994. CO-OCCURRENCE VECTORS FROM CORPORA VS. DISTANCE VECTORS FROM DICTIONARIES. In *Proc. COLING 94*, Vol. 1, pp. 304–309.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 1, pp. 448–453.

Yiming Yang and Christopher G. Chute. 1994. An Example-Based Mapping Method for Text Categorization and Retrieval. *ACM Transactions on Information Systems*, Vol. 12, No. 3, pp. 252–277.