Construction of a Simple Inference System of Textural Similarity (oka1 RITE-2)

Hirotaka Morita
Okayama University, Japan
morita@cl.cs.okayama-u.ac.jp

Koichi Takeuchi
Okayama University, Japan
koichi@cl.cs.okayama-u.ac.jp

ABSTRACT

The motivation to join the RITE-2 task is to understand what kinds of linguistic and common knowledge are needed to recognize textual similarity in RITE-2 data. After a shallow manual analysis of RITE-2-bc example data, we found that morpheme similarity would be a major factor to recognize textual similarity. Thus we construct a threshold based inference system with Japanese WordNet as a base system that can be extended to more deep linguistic knowledge in further development. The preliminary experimental results show that the simple noun similarity based system outperformed the WordNet based system; but in the formal run, the WordNet based system gives the highest score in f-measure among the other simple system we constructed.

Team Name

oka1

Subtasks

BC

Keywords

Morpheme-based Similarity, WordNet, Threshold

1. INTRODUCTION

In our laboratory, developing top-level ontology of predicate semantics and thesaurus of Japanese verbs [2], thus the basic motivation of submitting RITE-2 task was to apply our predicate semantic dictionary how to improve textual entailment recognition performance. When looking at the entailment examples in RITE-2, we found that the most of key issues to solve the similarity between sentences will be noun relations (see Section 2). Thus in this task we concentrate on constructing the base system of evaluating sentence similarities using WordNet toward applying more complex predicate semantics to the entailment texts in futue work.

In this manuscript, firstly we report the simple analyses of the entailment patterns in RITE-2 bc sample data, and then the constructed WordNet baed textual entailment system. After we show the applied results to bc of RITE and RITE2 bc mc data, we conclude this work.

2. PRELIMINARY ANALYSES OF ENTAILMENT PAIRS

Table 1: Rough sketch of entailment type in RITE2-bc

Type and examples	Pair id
Orthography and simple はがした/剥がした (peeled off) トータルファミリーブランド/ブランド total-family-brand/brand	9, 16, 29
Predicate alternations as 販売/発売 (sell/on sale) or 言う/差す (be expressed/indicate)	7, 23, 26
Complex construction with nouns X は Y の頭字語〜Z の一種/ Z の略称として〜頭字語 X〜 (X is an acronym of Y .. a kind of Z/ as an abbreviation of Z . . an acronym X) X 以外で規制/X で行われている (be prohibited except for X/ be done in X) X の原因仮説は Y/not (X は Y(の結果)) (causal hypothesis of X is Y/ not (X is the result of Y))	2, 4, 6, 11, 10, 12, 15, 20, 22, 24, 28
Extended semantics with human judgment 張り付けにされた/罰を受ける (be put on the cross/punishment)	1, 22
Knowledge of daily living もう一本もらえる/あたり付き (can get another one/ with winning number)	5

When we look at the top 30 examples in RITE-2-bc data, we found that the most of the essence to judge entailment or not are complex construction with noun-related argument structure: for example, “X, known for X’s Y which became a candidate of the Naoki prize” entails and “Y, the novel by X, was a candidate of the Naoki prize”. The rough sketch of the breakdown list is described in Table 1[1].

As shown in Table 1, the predicate alternations that are what our laboratory focus on would be a small part of feature to capture the entailment relations. Thus we decide to take into account similarities of nominal words between t1 and t2 as a first step to capture the complex nominal argument structure that must be the key issue to capture the entailment relations. The detailed descriptions of our

[1]No pair id in Table 1 indicates not entailment relation.

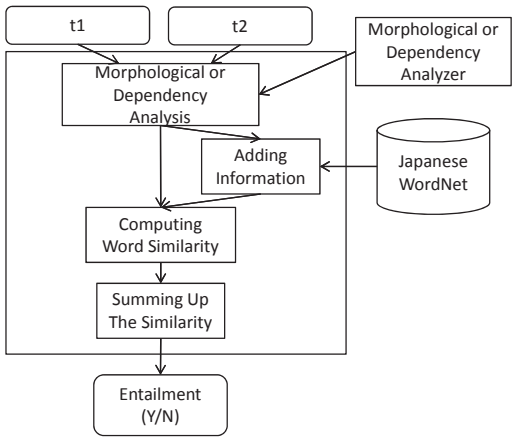

Figure 1: Overview of the proposed inference system of textual similarity

proposed system are below.

3. PROPOSED ENTAILMENT INFERENCE SYSTEM

Before applying some statistical learning approaches such as CRFs and SVMs, we try to understand what features are really effective for recognition of entailment relations. Thus we take a simple threshold approach for a targeting feature. The thresholds are manually decided using not only RITE-2-bc and RITE-2-mc sample data but also RITE1-bc and RITE-mc data in NTCIR-9. As natural language processing tools, we apply a morphological analyzer ChaSen[2] and a dependency analyzer CaboCha[3]. In preliminary experiments we found that ChaSen must be more preferable than Mecab to recognize morphemes. This is because Mecab has a tendency to recognize unknown words as proper nouns; but proper noun is one of the feature of discriminating entailment, and then noisy proper nouns must be unfavourable for the task (see also the results in Section 4.2). As a language resource to evaluate similarity between words, we apply Japanese WordNet [1].

The overview of the proposed entailment recognition system is depicted in Figure 1.

Every module is independent and can be freely added or removed, then we will apply three methods of evaluating similarity: (1) noun morpheme similarity, (2) chunk similarity, and (3) content morpheme similarity with WordNet. In the followings we describe the details of each method.

3.1 Content morpheme similarity

The ratio of the same content morphemes between t1 and t2 is applied to discrimination of entailment relation of them. The reason we take this feature comes from the following analysis of RITE-2-bc data. Table 2 shows that the number of the same morphemes between t1 and t2 in RITE-2-bc examples analyzed by MeCab and ChaSen.

In Table 2 the symbols P, GN, PD, CM AUX, VN, NSF, LOC, NUM, AL and FN indicate particle, general noun, period, comma, auxiliary verb, verbal noun, nominal suffix,

[2]http://chasen-legacy.sourceforge.jp/.

[3]http://code.google.com/p/cabocha/.

Table 2: Number of the POSes that the same morphemes between t1 and t2 in RITE2-bc sample data

MeCab		ChaSen	
POS	freq.	POS	freq.
助詞 (P)	1859	助詞 (P)	1840
名詞-一般 (GN)	1546	名詞-一般 (GN)	1416
記号-句点 (PD)	598	記号-句点 (PD)	599
記号-読点 (CM)	526	記号-読点 (CM)	525
助動詞 (AUX)	458	助動詞 (AUX)	470
名詞-サ変接続 (VN)	399	名詞-サ変接続 (VN)	382
名詞-接尾-一般 (NSF)	257	名詞-数 (NUM)	290
名詞-固有名詞-地域-一般 (LOC)	173	名詞-接尾-一般 (NSF)	241
名詞-数 (Num)	127	動詞 (V)	366
名詞-固有名詞-人名-姓 (FN)	112	記号-アルファベット (AL)	197
:	:	:	:
total	7294		7653

location, number, and family name, respectively. Comparing with the total matching number of Mecab and ChaSen, ChaSen can detect the same morphemes more than Mecab even though they are using the same dictionary; then we apply ChaSen to the task (see also Section 4.2).

To see how the ratio of the same morphemes is effective in entailment, the relations between POSes and entailment are visualize using a statistical learning tool Weka[4] in Figure 2. Each of the graph shows correlation between overlapping POS and Y/N label; the horizontal and vertical axes and show the ratio of the overlapping morpheme rates in a POS and the number of entailment pairs, respectively. The blue and red color parts show Y and N in entailments, respectively. In the upper graphs, the left side graph shows the correlations of nouns and the right side graph shows the correlations of verbs. In the lower graphs, the left and right side graphs are the correlations of proper nouns and symbols.

Comparing with the four POSes, the correlation of nouns shows a clear tendency that the number entailment Y pairs increase linearly with the ratio of the same morphemes. From the analyses, the ratio of overlapping morphemes in nouns are applied to discriminating entailments.

$$\text{Content morpheme similarity} = \frac{\#\text{ Overlapping morphemes of nouns in t1 and t2}}{\#\text{ Morphemes in t2}}$$

As the discrimination system, the system recognizes that the pair is entailment if the ratio of overlapping morphemes of nouns for the pair is over the threshold, and not entailment if the ration is under the threshold.

3.2 Chunk similarity

The second entailment discrimination system is based on similarity of chunks between t1 and t2. Chunk similarity indicates the ratio of the same chunks by the total chunks. The same chunk is identical chunks between t1 and t2 that the both chunks contain the same content words in the same

[4]http://www.cs.waikato.ac.nz/ml/weka/.

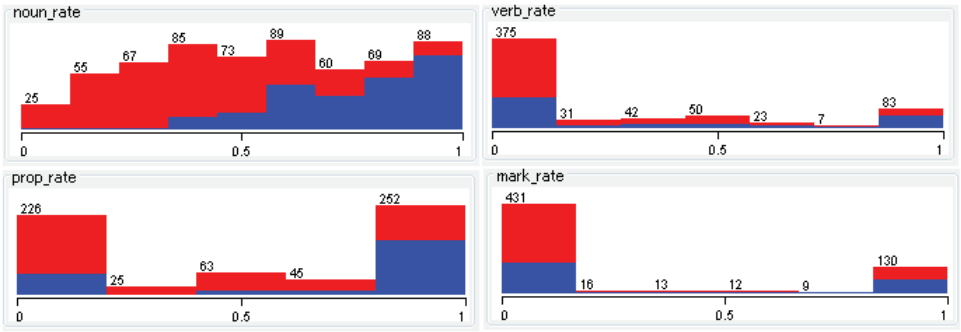

Figure 2: Correlation between ratio of overlapping POS and Y/N label in Weka analysis

morpheme order.

$$\text{Chunk similarity} = \frac{\#\text{ Same chunks in t1 and t2}}{\#\text{ Chunks in t2}}$$

As the discrimination system, the system recognizes that the pair is entailment if the ratio of chunk similarity for the pair is over the threshold, and not entailment if the ration is under the threshold.

3.3 Morpheme similarity with WordNet

An language resource Japanese WordNet is applied for taking into account the similarity between content morphemes. The content morphemes indicate noun, verb, adjective and adverbs[5]. Since all of the morphemes are located in the tree hierarchy of the synset, the similarity of two morphemes is calculated using the deepest common upper node of the both morphemes. Assuming that the depths of the tree for morphemes x and y are d_x and d_y and the depth of the deepest common upper node of the both morpheme is c_{xy}, we define the morpheme similarity below.

$$\text{Similarity with WordNet} = \frac{2c_{xy}}{(d_x+1)+(d_y+1)}$$

If a pair of morphemes does not exist in WordNet, the similarity is not 0 but 0.1 because we assume that unknown morpheme pairs still have a possibility of some similarity. Besides, if both morphemes are identical we give the similarity of 1.0 because of reducing the time of accessing to the tree of WordNet. Since a morpheme is sometimes registered in several nodes, then we take the highest value of similarity among the pairs of tree nodes in WordNet.

As the discrimination system, the system recognizes that the pair is entailment if the morpheme similarity with WordNet for the pair is over the threshold, and not entailment if the ration is under the threshold.

4. ENTAIL RECOGNITION EXPERIMENTS

Firstly we find the best threshold using the training corpora in RITE and RITE2 as development corpora, and finally we applied the proposed three discrimiantion systems to the test corpus in formal run.

4.1 Experimental setup

[5]We exclude numbers and functional words in nouns.

The Table3 shows the statistics of the training corpus distributed in RITE and RITE2. The symbol *bc* and *mc2bc* denote binary task data and binary task data converted from multi-classification data, respectively.

Table 3: Statistics of the development corpus

Data	#Y	#N	Total
RITE1-bc	250	250	500
RITE1-mc2bc	185	255	440
RITE2-bc	240	371	611
RITE2-mc2bc	290	258	548

We apply precision rates, recall rates and f-measure of correctly recognized Y or N to evaluation of performance. The equations are below.

$$\text{Precision of Y(N)} = \frac{\#\text{Correctly recognized as Y(N)}}{\#\text{Outputs of system as Y(N)}} \quad (1)$$

$$\text{Recall of Y(N)} = \frac{\#\text{Correctly recognized as Y(N)}}{\#\text{ Y(N) in the gold standard}} \quad (2)$$

$$\text{F of Y(N)} = \frac{2 \times \text{Precision Y(N)} \times \text{Recall Y(N)}}{\text{Precision Y(N)} + \text{Recall Y(N)}} \quad (3)$$

4.2 Experimental results for development corpus

Table 4 shows the best performance of the content morpheme similarity method on the development corpora. We found the best threshold value for macroF1 by shifting the threshold values from 0 to 1.0 by 0.01 step. Figure 3 shows the correlations between macroF1 and threshold. The characteristics of Figure 3 is that the macroF1 scores of all the data show the same tendency to the threshold value. The top values of macroF1 are around threshold = 0.6. This indicates that all of the entailment ratio of Y/N must be the same corresponding with the ratio of the same morphemes between t1 and t2.

How the noun morphemes are effective comparing with the other morphemes? To see this, we also apply the overlapping method to all morphemes. Figure 4 shows macroF1 scores of these two methods. According to the figure, the noun morpheme method obviously outperformed the all morpheme method; from the results the noun morpheme is a key factor to discriminate entailment relations. Besides we also apply

Table 4: Noun morpheme similarity

Data	Y/N	Prec.	Rec.	Accuracy	F1	MacroF1
R1	Y	52.96	75.20	54.20	62.15	52.09
bc	N	57.24	33.20	(271/500)	43.03	th: 0.63
R1	Y	64.85	70.81	71.59	67.70	71.17
mc2bc	N	77.31	72.16	(315/440)	74.65	th: 0.59
R2	Y	71.48	78.33	79.21	74.75	78.54
bc	N	85.06	79.78	(484/611)	82.34	th: 0.64
R2	Y	77.42	82.76	78.10	80.00	77.90
mc2bc	N	78.99	72.87	(428/548)	75.81	th: 0.59

Table 5: Chunk similarity

Data	Y/N	Prec.	Rec.	Accuracy	F1	MacroF1
R1	Y	52.32	76.80	53.40	62.24	50.70
bc	N	56.39	30.00	(267/500)	39.16	th: 0.25
R1	Y	55.67	61.08	63.18	58.25	62.66
mc2bc	N	69.62	64.71	(278/440)	67.05	th: 0.3
R2	Y	64.68	67.92	72.83	66.26	71.76
bc	N	78.55	76.01	(445/611)	77.26	th: 0.31
R2	Y	71.13	69.66	68.98	70.38	68.91
mc2bc	N	66.67	68.22	(378/548)	67.43	th: 0.29

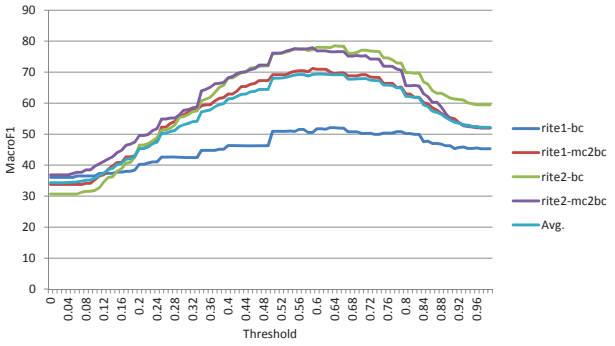

Figure 3: Correlations between threshold and macroF1 on content morpheme similarity method

ChaSen (denoted as 'c') and Mecab (as 'm') for looking at the differences between morphological analyzers.

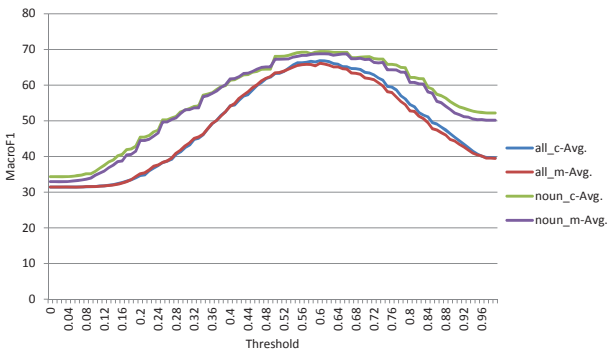

Figure 4: Overlapping method with noun vs. all morphemes / ChaSen vs. Mecab

Table 4 shows that the system using ChaSen outperformed that using Mecab on all of the threshold values. Thus we use ChaSen for the following systems.

Table 5 shows the results of chunk similarity method, and Figure 5 shows the correlations between macroF1 and threshold for chunk similarity.

The characteristics of Figure 5 is that the macroF1 scores of all the data show the same tendency to the threshold value. The top values of macroF1 are around threshold = 0.3. Comparing with the performance of the system using morpheme similarity in Figure 3, the top values of macroF1 of chunk similarity decrease from 1 to 8 scores in macroF1 than that of morpheme similarity. Thus we apply morpheme similarity to the revised entail recognition system described in the next paragraph.

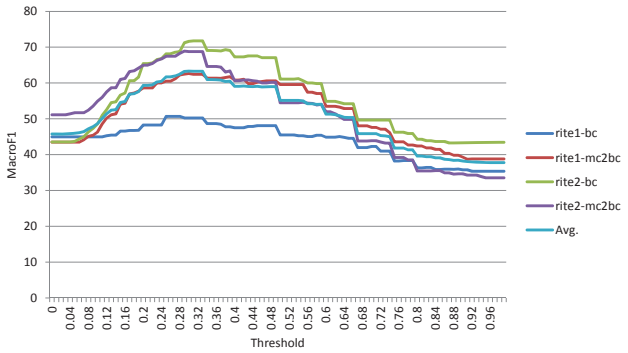

Figure 5: Correlations between threshold and macroF1 on chunk similarity method

Table 6 shows the results of morpheme similarity with WordNet. The thereshold is fixed as 0.77; this figure is determined by the analyses of WeKa on both RITE2-bc (Figure 6) and RITE2-mc2bc (7). The both figures reveal that if threshold value is over than 0.77, the number of entailment Y pair exceeds the number of N. We did not test all of the macroF1 values changing the threshold value by 0.01 step since it needs a lot of time due to time-consuming process of accessing to WordNet data.

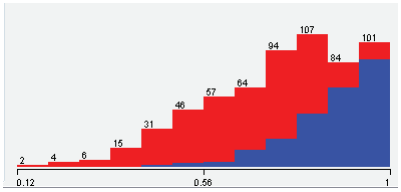

Figure 6: Correlation between similarity with WordNet and Y/N label on bc corpus in Weka analyses

Comparing Table 6 with Table 4 and Table 5, we found that macroF1 score of WordNet based system is lower than that of morpheme based system, but it is higher than that of chunk based system except for RITE-bc data.

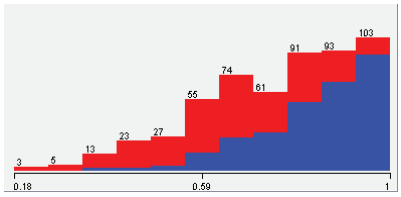

Figure 7: Correlation between similarity with WordNet and Y/N label on mc2b corpus in Weka analyses

When we apply the above three methods to the test corpus in formal run, all of the proposed methods outperformed the baseline system in macroF1 scores (see Table 7)[3]. Comparing with the proposed methods, the WordNet based system performed the best results of all. Table 7 shows that the noun based system detect Y/N relations the balance of detecting Y/N relations are biased, while the WordNet based system detect Y/N not biased a lot; thus the F score of the WordNet system is the best score. This indicates that WordNet based system is robust for the entailment recognition task.

Table 6: Morpheme similarity with WordNet

Data	Y/N	Prec.	Rec.	Accuracy	F1	MacroF1
R1	Y	49.68	62.80	49.60	55.48	48.71
bc	N	49.46	36.40	(248/500)	41.94	
R1	Y	65.60	44.32	66.82	52.90	63.64
mc2bc	N	67.30	83.14	(294/440)	74.39	
R2	Y	69.52	77.92	77.91	73.48	77.27
bc	N	84.50	77.90	(476/611)	81.07	
R2	Y	76.21	70.69	72.81	73.35	72.80
mc2bc	N	69.53	75.19	(399/548)	72.80	

Table 7: Results in formal run

Methods	Y/N	Prec.	Rec.	F1	MacroF1
Noun	Y	64.55	87.50	74.30	74.59
	N	87.83	65.25	74.88	th: 0.6
Chunk	Y	65.36	71.48	68.28	71.71
	N	77.88	72.60	75.15	th: 0.3
WN	Y	70.71	77.34	73.88	76.71
	N	82.42	76.84	79.53	th:0.77
Base	Y	57.63	53.13	55.28	62.53
	N	67.91	71.75	69.78	

5. CONCLUSIONS

We construct a threshold-based inference system with Japanese WordNet as a base system that can be extended to more deep linguistic knowledge in further development. The preliminary experimental results show that the simple morpheme similarity based system outperformed the WordNet based system; but in the formal run, the WordNet-based system gives the highest score 76.71 in f-measure among the other simple system we constructed. In future work we will apply more deep linguistic knowledge to entailment recognition system.

6. REFERENCES

[1] F. Bond, H. Isahara, K. Kanzaki, and K. Uchimoto. Construction of Japanese WordNet from Multi-lingual WordNet. In *Proceedings of the 14th Annual Meeting of Japanese Natural Language Processing*, pages 853–856, 2008.

[2] K. Takeuchi, K. Inui, N. Takeuchi, and A. Fujita. A thesaurus of predicate-argument structure for japanese verbs to deal with granularity of verb meanings. In *The 8th Workshop on Asian Language Resources*, pages 1–8, 2010.

[3] Y. Watanabe, Y. Miyao, J. Mizuno, T. Shibata, H. Kanayama, C.-W. Lee, C.-J. Lin, S. Shi, T. Mitamura, N. Kando, H. Shima, and K. Takeda. Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. In *Proceedings of the 10th NTCIR Conference*, 2013.