UIOWA at NTCIR-9 RITE: Using the Power of the Crowd to Establish Inference Rules

Christopher G. Harris
Informatics Program
The University of Iowa
Iowa City, IA 52242

christopher-harris@uiowa.edu

ABSTRACT
We participated in the Binary Classification (BC), Multiple Classification (MC), and Question and Answer (RITE4QA) subtasks for both Simplified Chinese and Traditional Chinese in NTCIR-9 RITE. In this paper, we describe our procedure to establish inference rules using crowdsourcing, refine and weigh them, and apply these rules to a test collection.

Categories and Subject Descriptors
H.3.3 Information Storage and Retrieval – Retrieval Models

General Terms
Algorithms, Performance, Design, Experimentation

Keywords
Inference Models, Natural Language Processing, Crowdsourcing, Human Computation, Information Access

1. INTRODUCTION
This was UIOWA's first year to participate in the RITE (Recognizing Inference in Text). This task consisted of evaluating a pair of text segments (in the form of sentences) and determining the entailment (inference) between the pairs. In the Binary Classification subtask, the decision was binary (yes or no) to establish entailment between the text pairs; for the Multiple Classification subtask, there was a five-way decision to be made in terms of direction of entailment. The RITE task provided Japanese, Simplified Chinese and Traditional Chinese subtasks – UIOWA participated in the Simplified and Traditional Chinese subtasks, and the RITE4QA Question and Answer task. Detailed information about the NTCIR-9 RITE task can be found in the task overview [10].

This paper is organized as follows. In the next section, we describe the overall experimental system we implemented for this task, including the rules and procedures we incorporated in our submitted runs. In Section 3 we present the experimental results for our runs and discuss how some of the components incorporated affected the results. We conclude our discussion in Section 4.

2. SYSTEM DESCRIPTION
With no previous experience in evaluating entailment in text pairs and little knowledge of either Simplified or Traditional Chinese text, we undertook this challenge primarily to test out the establishment of NLP methods using crowdsourcing to develop and help test the rule sets. Our approach is based on the knowledge that humans approach their understanding of entailment based on established language rules (which we call rule sets), and thus are the best resource to provide them. Computers are better suited to enforce these established rules - once the rules have been identified. Thus the challenges are to identify as many rules as feasible, and when these rules conflict or overlap, decide which of the rules should have the greatest influence on our entailment decisions.

Each participating RITE team was provided with training data for both BC and MC evaluations for all language subtasks. Our first process was to parse the Chinese text and identifying and tag the parts of speech.

2.1 Parsing and Tokenization
We used the Stanford Parser [5, 9] to tokenize, determine token dependencies, and perform part of speech (POS) tagging. The Stanford Parser relies on a linear-chain conditional random filed (CRF) model, which treats word segmentation as a binary decision task. The tool makes use of features such as character identity n-grams, morphological features and character reduplication features. The word segmentation tool exploits lexicons and proper noun features to improve segmentation consistency.

Unfortunately, the Stanford Parser's native ability to parse our text segments was insufficient in both traditional and simplified Chinese, so we manually developed a modified Chinese text segmentation tool using lingpipe [3] to fill in the parsing inconsistencies. This segmentation tool was used to segment the Chinese sentences into appropriate tokens. The most frequent issues were with named entities. Examining and properly tokenizing the sentences required a fair amount of effort; however, we believe the tool is a necessary step in properly identifying parts of speech in each sentence.

Our approach depends heavily on determining the relationships between our identified tokens; to address this task of relationship identification, we manually created a thesaurus-like set of synonyms developed from corpus terms. Our approach used several different resources, including a CC-CEDICT-based dictionary tool [4] and some Chinese WordNet tools [6, 7]. We were able to produce a thesaurus of established synonyms for these tokens. We applied a part-of-speech tagger developed in-house for tokens identified by lingpipe but missed or mishandled by the Stanford Parser.

The Appendix of this paper provides two examples of our parsing system taken from entailments in the test data.

Due to the frequency of geographic terms in the dataset, we incorporated the use of additional geographic information as synonyms to expand our queries of identified geographic terms. To accomplish this, we first separated and identified the known geographic tokens in the dataset. We then incorporated the use of 5.9 million geographic terms obtained from the Alexandria Digital Library[1][1]. Next, we translated these terms into Chinese using the CC-CEDICT dictionary and used this information to establish relationships between geographic terms that existed between our tokens.

With the tokens, dependencies, and their parts of speech identified, and a thesaurus established, we turned to the crowd to help us derive specific rules that could be applied to entailment decisions in our training set. Through several crowdsourcing platforms, we recruited native Chinese speakers familiar with Traditional Chinese and/or Simplified Chinese. We provided them with a randomly-selected portion of text pairs consisting of two-thirds of each training set (we held one-third of the provided training set as our tuning set, see Table 1 for how the test/tuning sets were divided).

Table 1. Description of the RITE Tuning and Training Data

Subtask	Training Data Pairs Provided by RITE Organizers*	Data provided to the Crowd for Rule-building	Data Withheld as a Tuning Set
CT-BC	421 Pairs	280 Pairs	141 Pairs
CT-MC	421 Pairs	280 Pairs	141 Pairs
CS-BC	407 Pairs	270 Pairs	137 Pairs
CS-MC	407 Pairs	270 Pairs	137 Pairs

We provided the crowd with the original entailment, the segmentation and part of speech tags, dependencies, and the synonyms identified. We requested for them to provide us with rules based on these components. We gave them the same set of instructions provided to us by the RITE organizers, but instead of asking them to score each pair, we asked them to *report the rules they used* for scoring each pair. This initial crowdsourcing component took about one week to complete. Demographic information about each of these participants is provided in Section 2.2.

Below is the information provided to the crowd for one such text pair from the test set:

2.1.1 Traditional Chinese Text Given

t_1: **丁磊石网易的**创办**人**

t_2: **丁磊**1997**年**6**月創立網易公司**

2.1.2 English Translations

t_1: NetEase's founder Ding Lei Shi

t_2: Ding established the company NetEase in June 1997

2.1.3 Segmentation and Tagging of t_1

A1. 丁磊石/NN
A2. 网易/NN
A3. 的/DEG
A4. 创办人/NN

2.1.4 Typed dependencies (collapsed) for t_1

nn(创办人- A4, 丁磊- A1)
assmod(创办人- A4, 网易- A2)
assm(网易- A2, 的- A3)

2.1.5 Synonyms for Tokens in t_1

A1 丁磊石: 丁; 磊石
A2 网易 : 公司 ; 站 ; 有限公司 ; 企業 ; 廠商
A4 创办人: 创立者; 缔造者; 创立者

2.1.6 Segmentation and Tagging for t_2

B1. 丁磊/NR
B2. 1997/CD
B3. 年/M
B4. 6/CD
B5. 月/NN
B6. 創立/NN
B7. 網易/NN
B8. 公司/NN

2.1.7 Typed dependencies (collapsed) for t_2

nn(創立- B6, 丁磊- B1)
nummod(年- B3, 1997- B2)
clf(創立- B6, 年- B3)
nummod(月- B5, 6- B4)
nn(創立- B6, 月- B5)
nn(公司- B8, 創立- B6)
nn(公司- B8, 網易- B7)

2.1.8 Synonyms for Tokens in t_2

B1 丁: 丁磊石; 磊石
B2 1997: 97; 一九九七; 九七; 丁丑
B3 年: 歲; 載; 茲; 年下; 年表 ; 學年 ; 周歲
B4 6: 六 : 六月份 ; 第六
B5 月: 逐月; 月亮; 月食; 满月; 残月; 月初; 月末, 月报
B6 創立: 起 ; 建 ; 設 ; 建立 ; 創辦 ; 成立 ; 設立 ; 開闢 ; 創建 ; 建成 ; 立下 ; 奠定 ; 確立 ; 樹立 ; 植根 ; 創設 ; 建樹 ; 創牌子; 创办
B7 网易 : 公司 ; 站 ; 有限公司 ; 企業 ; 廠商
B8 公司: 站 ; 有限公司 ; 企業 ; 廠商; 株式会社

[1] http://alexandria.ucsb.edu/gazetteer/

The following is a subset of the 21 different rules submitted by the crowd that would apply to the above information:

R12: Synonyms between sentences
R19: Same objects in both sentences
R24: A noun in one sentence a synonym for another proper noun in the other sentence
R30: Additional information in one of the sentences tied to the same synonym
R33: One sentence longer than another
R37: One sentence with dates

The crowd provided additional information about the entailment based on these rules (e.g., "R12: More than one synonym between sentences *usually means that it will be a B, F, or R*"), but we removed this additional information. Instead, we used SVM^{light} to determine the most likely class based on a series of four binary decisions (see Figure 1) based the predicate of the submitted rule (e.g., "More than one synonym between sentences"). SVM^{light} allowed us to examine each of the four binary decisions in Figure 1 based on the synonyms in the two sentences. Information about our use of SVM^{light} is examined further in Section 2.3.

Even with a rich set of submitted rules, many submitted rules needed to be refined so they could be machine-interpreted. For example one submitted rule stated: "If we have two sentences, and one sentence is more positive than another, the entailment will likely favor the positive sentence" would not be helpful if we are unable to adequately determine the 'degree of positivity,' or polarity, of each of the sentences. We were required to rewrite the rule to refer to measurements the system could establish, such as number of adjectives describing the subject, word counts, number of matching synonyms between the sentences, etc. For example, the predicate for rule R12 could be established as:

$$\text{count}\{\text{syn}(t_1),\ \text{syn}(t_2)\} > 1$$

Likewise, the predicate for the rule R33 could be established as:

$$\text{len}\{(t_1) > (t_2)\}$$

This is favorable, since it is now a simple binary decision to examine. Table 2 illustrates some of the most common rule types established by the crowd based on the test data provided.

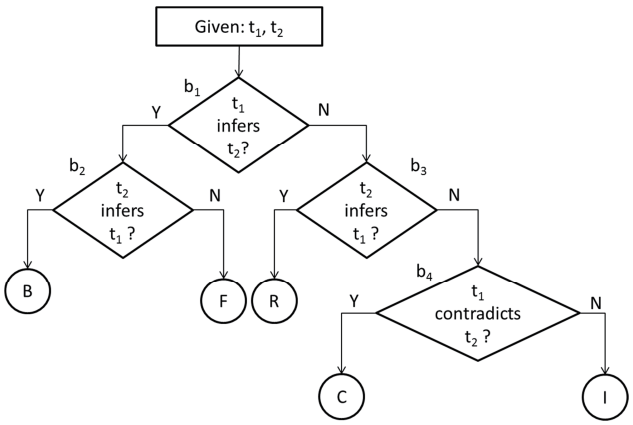

Figure 1. Flowchart of the four binary decisions b_1 to b_4 (represented as diamonds) in the Multiple Class (MC) subtask.

Note that the decisions in the binary (BC and RITE4QA) subtasks followed the same rule creation process, but only involved a single entailment decision ('Y' or 'N').

2.2 Crowdsourcing Participants

Using three crowdsourcing websites (eLance[2], Guru[3], and oDesk[4]), we posted our request for native Chinese speakers to provide us with rules for the data, with the specific instruction that only commonly-known knowledge (i.e., no outside research) was permitted. We received 16 applicants, nine understood both Traditional and Simplified Chinese scripts, two claimed reading knowledge of Traditional Chinese only, and five claimed reading knowledge the Simplified Chinese script only. Six participants did not submit rules that were usable; ten participants submitted rules which we incorporated (three supplied rules for Traditional Chinese, three of provided rules for Simplified Chinese and four provided for both).

Of these ten participants, eight had at least one year of university education; two were located in Taiwan, two were in Hong Kong, one was in Malaysia, and five specified mainland China as their location. We paid some participants by the hour, some as a fixed fee, and some by the number of valid rules they supplied. There was a wide variety of payments made to our participants, and we will analyze and report how the quality of the information provided was affected by the payment methods and amounts in a separate study.

Table 2. Top Five Rules Types for Traditional and Simplified Chinese subtasks

Rule Type	Percentage of all Rule Submissions
Simple aggregates (word counts, counts of nouns and noun phrases, number of prepositions, etc.)	13.1%
Synonym matching between tokens in the two sentences	12.6%
Subtle differences between verbs or adjectives in the two sentences	10.4%
Clarification of named entities (e.g., use of 那个or 哪一个)	9.1%
Additional explanation or clarification in one sentence as compared with a second sentence	8.2%

2.3 Rule Set Resolution

From our crowdsourcing participants, we received a total of 78 rules in Traditional Chinese and 103 in Simplified Chinese. Many of the rules in each script could also be applied to the other (e.g. many of the rules for Simplified Chinese term pairs could also apply to Traditional Chinese term pairs). Some of the rules were duplicates and were removed. Based only on the test data provided, we ended up with 73 useable rules in Traditional

[2] http://www.elance.com
[3] http://www.guru.com
[4] http://www.odesk.com

Chinese and 69 in Simplified Chinese. We then implemented mex-svm[5], a Matlab interface for SVM^{light} [7] on the training dataset by examining each of the four binary decisions (the diamonds shown in Figure 1).

SVM^{light} incorporated all of the submitted rules to determine the relative importance of each rule for the entailment classification for each of the four binary decisions (b_1 to b_4). Evaluating the results for each of these binary decisions separately allowed us to see which of the submitted rules were most essential for determining entailment. We also compared the expectations from the crowd (recall that the submission by the crowd for R12 was "More than one synonym between sentences *usually means that it will be a B, F, or R*"). We would compare the entailment decision from SVM^{light} for that predicate to see if the rule alone actually obtained a "B", "F", or "R"; if these classes were not the most likely to appear, we would investigate why.

Unfortunately the results from this initial run demonstrated were far from satisfactory; once again we solicited assistance from the crowd to help us fill in the gaps and find issues with the rules that had already submitted.

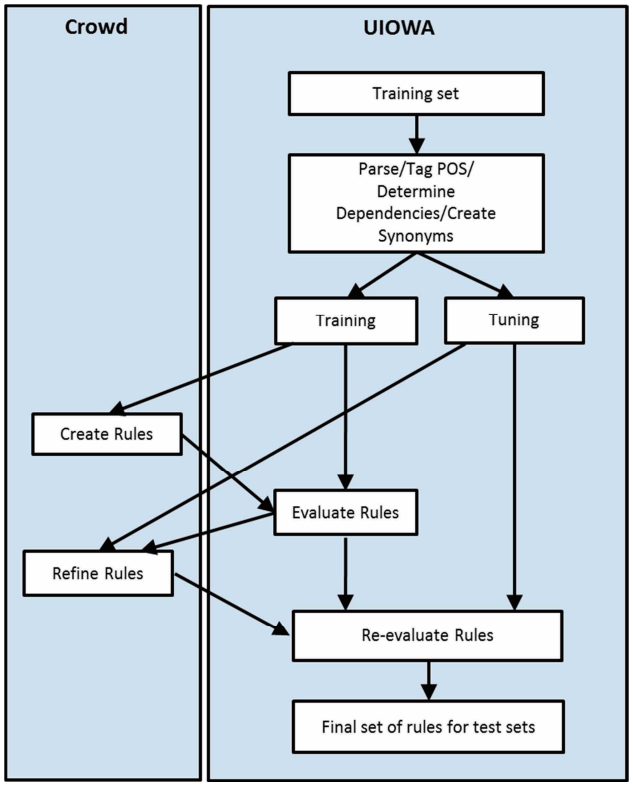

Figure 2. Flow of information between processes, showing the division of tasks between the crowdsourcing participants and UIOWA.

2.4 Additional Refinement

We provided each crowdsourcing participant with the full set of rules we determined for each of the languages, as well as the tuning set originally withheld from them when determining their initial set of rules. We then asked the crowd to examine these previously-submitted rules, the results we obtained from our initial run, and suggest any refinements to the previously-submitted rules.

We had some participants provide us with new rules at this stage. Six of the participants supplied us with at least one additional rule, and a total of 7 new rules in Traditional Chinese and 8 in Simplified Chinese were obtained through this additional refinement step.

In a few of our binary (BC) entailment runs, we noticed that SVM^{light} did quite well for those that scored far from zero, but as we got closer to zero, the ability for SVM^{light} to distinguish between the "Y" and "N" classes was less satisfactory, so we manually tweaked the rule sets to favor one class over the other in these marginal cases. Those rule sets favoring the positive (Y) class are considered tight, or restrictive, whereas those that favor the negative class are considered relaxed or flexible.

Table 3. Description of UIOWA's RITE runs

Run	Description
CT-BC-1	Used refined crowdsourcing-developed rules with a slight bias towards 'yes'
CT-BC-2	Used refined crowdsourcing-developed rules with a slight bias towards 'no
CT-BC-3	Used only the refined crowdsourcing-developed rules that a single participant submitting the most rules agreed with
CT-MC-1	Used refined crowdsourcing-developed rules with a tight (restrictive) bias
CT-MC-2	Used refined crowdsourcing-developed rules with a relaxed (flexible) bias
CT-MC-3	Used only the refined crowdsourcing-developed rules that a single participant submitting the most rules agreed with
CS-BC-1	Used refined crowdsourcing-developed rules with no bias
CS-BC-2	Used only the refined crowdsourcing-developed rules that a single participant submitting the most rules agreed with
CS-MC-1	Used refined crowdsourcing-developed rules with a tight (restrictive) bias
CS-MC-2	Used refined crowdsourcing-developed with a relaxed (flexible) bias
CS-MC-3	Used only the refined crowdsourcing-developed rules that a single participant submitting the most rules agreed with
CT-RITE4QA	Used refined crowdsourcing-developed rules with a slight bias towards 'yes'
CS-RITE4QA	Used refined crowdsourcing-developed rules with a slight bias towards 'yes'

3. RESULTS

3.1 Submitted Results

Each task scored each pair as correct or incorrect, allowing this binary decision to provide a score indicating accuracy. We provide the results for each run as well as the subtask average in Table 4.

[5] http://sourceforge.net/projects/mex-svm/

Our performance is relatively good compared to the average RITE submission (all of UIOWA submitted runs were higher than the average score). From this information, we do notice some interesting results. For Simplified Chinese, although the MC has five possible values and BC only two, the scores are almost identical between these two for all participants including UIOWA; for Traditional Chinese, the disparity between these scores is much larger, not only in our scores but other teams as well. This may point to a more challenging dataset used with Simplified Chinese.

Table 4. Scoring for Submitted UIOWA RITE Runs Comparing to the NTCIR-9 RITE Task Average

Run	Score	Task Average
CT-BC-1	0.971	0.714
CT-BC-2	0.936	
CT-BC-3	0.963	
CT-MC-1	0.787	0.502
CT-MC-2	0.774	
CT-MC-3	0.724	
CS-BC-1	0.908	0.621
CS-BC-2	0.884	
CS-MC-1	0.892	0.597
CS-MC-2	0.892	
CS-MC-3	0.887	
CS-RITE4QA	0.901	*
CT-RITE4QA	0.901	*

* - this information was unavailable at publication time

We see that the 'tight' bias (Run 1 in both CT-MC and CS-MC) is preferable to the 'relaxed' bias (Run 2 in both CT-MC and CS-MC). Also, implementing a slight bias towards 'yes' in CT-BC subtask provided better results than a run biased towards 'no'. This may indicate that in a situation where the decision is close, better results can be obtained by a slight favoritism to the positive (Y) case rather than the negative (N) class. We do note that this bias is likely data collection dependent.

3.2 Comparison with Scores from Individual Participants

After this run was completed, we asked three participants from the crowd who did not contribute to the rules to make Binary and Multiple Classification judgments on the test dataset for both Traditional and Simplified Chinese subtasks. The results were not submitted, but are provided in Table 5. We compare these results to the best scored UIOWA run for each of the four subtasks.

We see that the scores obtained using the rules established by the crowd (our submitted runs) - a hybrid approach between human *rule creators* and machine *rule interpreters* - actually can score slightly higher than an individual taking the same test in some cases. This reinforces research in psychology which illustrates the power of group diversity over individuals in performing complex tasks, such as developing rules for an entailment task [2].

Examining the confusion matrix generated for the individual test takers, we see that the decisions between entailment for 'both' (B) and the 'forward' (F) and 'reverse' (R) entailment were the most difficult for individuals to correctly ascertain; however, with input from the crowd, the ability to make this determination was substantially improved.

Table 5. Scoring for the Best Submitted UIOWA RITE Runs Compared with the Highest Score from Crowdsourcing Participants

Run	Highest Crowdsourcing Participant Score	UIOWA Best Submitted Score
CT-BC	0.939	0.971 (Run 1)
CT-MC	0.823	0.787 (Run 1)
CS-BC	0.843	0.908 (Run 1)
CS-MC	0.838	0.892 (Run 1,2)

4. CONCLUSION

In NTCIR-9 RITE, we applied statistical parsing for both Traditional and Simplified Chinese text pairs to determine entailment. We applied a two-phase crowdsourcing approach to identify and examine language-specific rules and apply them to the text pairs. When compared to other machine-based approaches and human-based approaches, it appears that our hybrid approach can, in some cases, outperform a single human participant taking the same test.

We should note that our method relies heavily on the strength of good crowdsourcing participants to identify and re-evaluate rules. A substantial manual component was also involved in establishing and checking synonyms, refining crowd-submitted rule sets, etc. that we hope to more fully automate in future entailment examinations. Also, we make an assumption that the test and training data was randomly assigned, if they came from two very different collections, we believe our results would be negatively impacted.

5. REFERENCES

[1] Alexandria Digital Library Gazetteer. 1999- . Santa Barbara CA: Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. Copyright UC Regents. http://www.alexandria.ucsb.edu/gazetteer

[2] American Psychological Association. Groups Perform Better Than The Best Individuals At Solving Complex Problems. *Science Daily*. (2006, April 23). http://www.sciencedaily.com/releases/2006/04/060423191907.htm (Accessed September 14, 2011)

[3] Alias-i. LingPipe 4.1.0. http://alias-i.com/lingpipe 2008. (Accessed July 22, 2011)

[4] CC-CEDICT Chinese Word Dictionary. http://www.mdbg.net/chindict/chindict.php. 2011. (Accessed July 17, 2011).

[5] Chang, P., Tseng, H., Jurafsky, D., and Manning, C.D. 2009. Discriminative Reordering with Chinese Grammatical Relations Features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*. 2009.

[6] Huang C.R. and Chen K. J. Academic Sinica Balanced Corpus, Technical Report 95-02/98-04, Academic Sinica, Taipei, 1995.

[7] Huang C.R., S. K. Hsieh, J. F. Hong, Y. Z. Chen, I. L. Su, Y. X. Chen and S. W. Huang. 2008. Chinese Wordnet: Design, Implementation, and Application of an Infrastructure for Cross-lingual Knowledge Processing. *In Proc. of the 9th Chinese Lexical Semantics Workshop*. 2008

[8] Joachims, T. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[9] Levy, R. and Manning, C.D. 2003. Is it harder to parse Chinese, or the Chinese Treebank? *ACL 2003*, pp. 439-446.

[10] Shima, H., Kanayama, H., Lee, C.-W., Lin, C.-J., Mitamura, T., Miyao, Y. Shi, S., and Takeda, K. Overview of NTCIR-9 RITE: Recognizing Inference in TExt. *In NTCIR-9 Proceedings*, to appear, 2011.

6. APPENDIX

Below are two examples of determining the rules for entailment exercises. This provides an overview of the information presented to the crowd as well as the rules evaluated for each of the sentences. Although SVM^{light} evaluates all of the rules, including those not mentioned in the examples, to determine the appropriate class, we present those rules that are the most relevant. Note that some of the synonyms (particularly those based on IS-A-PART-OF relationships between geographic terms) are extensive and those not relevant to the example have not been included due to space considerations.

6.1 Example of Traditional Chinese Parsing

6.1.1 Traditional Chinese Text Given

t_1: 1997年香港回歸中國

t_2: 香港的主權和領土是在1997由英國歸還給中國的

6.1.2 English Translations

t_1: 1997 handover of Hong Kong to China

t_2: Hong Kong's sovereignty and territories in 1997 were returned to China by the British.

6.1.3 Segmentation and Tagging of t_1

A1. 1997/CD
A2. 年/M
A3. 香港/NR
A4. 回歸/NN
A5. 中國/NN

6.1.4 Typed dependencies (collapsed) for t_1

nummod(年-A2, 1997-A1)
clf(中國-A5, 年-A2)
nn(中國-A5, 香港-A3)
nn(中國-A5, 回歸-A4)

6.1.5 Synonyms for Selected Tokens in t_1

A1 1997: 97; 一九九七; 九七; 丁丑

A2 年: 歲; 載; 茲; 年下; 年表 ; 學年 ; 周歲

A3 香港： 大中華區； 港： 全港： 觀塘 屯門： 大埔： 金鐘； 特別行政區； 九龍城； 港澳 深水埗；兩岸三地；香港島； 港臺； 灣仔； 荃灣； 旺角；維港； 維多利亞港; 维多利亚港； 新界； 尖沙咀； 深港。。。

A4 回歸： 回； 回到；归还; 交還；返還；返回； 回來；退換貨；還給； 歸； 退還； 復； 回報；歸還給; 重回 ; 酬； 回敬； 歸還； 回還； 回返； 還禮； 奉還； 送還； 璧還。。。

A5 中國： 中； 華；中華民國； 中華； 大陸； 我國； 中華人民共和國；華夏；神州；中臺；大中華區。。。

6.1.6 Segmentation and Tagging for t_2

B1. 香港/NR
B2. 的/DEG
B3. 主權/NN
B4. 和/CC
B5. 領土/NN
B6. 是/VC
B7. 在/P
B8. 1997/CD
B9. 由/P
B10. 英國/NR
B11. 歸還給/VV
B12. 中國/NN
B13. 的/DEC

6.1.7 Typed dependencies (collapsed) for t_2

assmod(領土-B5, 香港-B1)
assm(香港-B1, 的-B2)
conj(領土-B5, 主權-B3)
cc(領土-B5, 和-B4)
top(是-B6, 領土-B5)
prep(歸還給-B11, 在-B7)
pobj(在-B7, 1997-B8)
prep(歸還給-B11, 由-B9)
pobj(由-B9, 英國-B10)
attr(是-B6, 歸還給-B11)
dobj(歸還給-B11, 中國-B12)
cpm(歸還給-B11, 的-B13)

6.1.8 Synonyms for Selected Tokens in t_2

B1 香港： 大中華區； 港： 全港： 觀塘 屯門： 大埔： 金鐘； 特別行政區； 九龍城； 港澳 深水埗；兩岸三地；香港島； 港臺； 灣仔； 荃灣； 旺角；維港； 維多利亞港; 维多利亚港； 新界； 尖沙咀； 深港。。。

B3 主權；統治權；至高統治權；王；治國；統治；法治；暴政；独裁;主宰；下轄；宰制；依法治國；按立憲治國；柄政；柄國

B5 領土：全港； 土地；地方；版圖；境； 領地； 疆域

；屬地；疆土；租借地
B8 1997: 97; 一九九七; 九七; 丁丑
B10 英國：英；英格蘭；不列顛；大不列顛；威爾士；蘇格蘭； 北愛；北愛爾蘭：倫敦。。。
B11 歸還給: 回歸; 回； 回到； 交還；返還；返回； 回來； 退換貨；還給； 歸； 退還； 復； 回報； 重回；酬; 回敬； 歸還； 回還； 回返； 還禮； 奉還； 送還；璧還。。。
B12 中國： 中； 華；中華民國； 中華； 大陸； 我國； 中華人民共和國；華夏；神州；中臺；大中華區。。。

6.1.9 Entailment Decision Logic
Some of the rules that influence the decision:

R12: A1 and B3 refer to a single synonymous NR.
R14: A4 and B11 are synonyms referring to 'a return of something'
R19: Dependent objects A5 and B12 are the same.
R33: One sentence longer than another
R37: One sentence with dates

Through our rule sets, a limited choice of "B" or "R" to describe the relationship. Additional information (B9, B10, B11) describes the relationship between t_1 and t_2 (t_2 entails t_1). Therefore, our rules establish "R" as the most likely answer. Choice "R" is the correct answer.

6.2 Example of Simplified Chinese Parsing

6.2.1 Simplified Chinese Text Given
t_1: 安南来自非洲加纳
t_2: 安南来自亚洲

6.2.2 English Translation
t_1: Annan from Africa and Ghana
t_2: Annan from Asia

6.2.3 Segmentation and Tagging of t_1
A1. 安南/NR
A2. 来自/VV
A3. 非洲/NR
A4. 加纳/NR

6.2.4 Typed dependencies (collapsed) of t_1
nsubj(来自-A2, 安南-A1)
nn(加纳-A4, 非洲-A3)
dobj(来自-A2, 加纳-A4)

6.2.5 Synonyms for Selected Tokens in t_1
A1 安南： 科菲安南；联合国秘书长； 安全理事会
A3 非洲：非； 撒哈拉以南非洲
A4 加纳：阿克拉

6.2.6 Segmentation and Tagging of t_2
B1. 安南/NR
B2. 来自/VV
B3. 亚洲/NR

6.2.7 Typed dependencies (collapsed) of t_2
nsubj(来自-B2, 安南-B1)
dobj(来自-B2, 亚洲-B3)

6.2.8 Synonyms for Selected Tokens in t_2
B1 安南： 科菲安南；联合国秘书长； 安全理事会
B3 亚： 亚洲； 欧亚； 亚太区； 大中华区； 亚太；泰东； 亚洲太平洋地区； 极东； 亚细亚；亚细亚洲；亚洲与太平洋地区。。。

6.2.9 Entailment Decision Logic
Some of the rules that influence the decision:

R12: A1 and B3 refer to a single synonymous NR
R3: verbs A2 and B2 are same VV.
R19: dependent objects A3 and A4 are not synonyms from B3.

Therefore, our choice is "C" as the relationship between t_1 and t_2 (t_1 and t_2 conflict with each other). Choice "C" is the correct answer.