LTI's Textual Entailment Recognizer System at NTCIR-9 RITE

Hideki Shima Yuanpeng Li Naoki Orii Teruko Mitamura

Language Technologies Institute, School of Computer Science, Carnegie Mellon University

5000 Forbes Ave, Pittsburgh, PA, USA

{hideki, yuanpeng.li, norii, teruko}@cs.cmu.edu

ABSTRACT

This paper describes the LTI's system participated in NTCIR-9 RITE. The system is based on multiple linguistically-motivated features and an adaptable framework for different datasets. The formal run scores are 54.6% (accuracy in BC), 66.7% (accuracy in Entrance Exam), and 29.8% (MRR in RITE4QA) which outperformed strong baselines, and are relatively good among participants. We also describe in-house experimental results (e.g. ablation study for measuring feature contribution).

Keywords: *textual entailment, near-synonym, domain adaptation*

1. INTRODUCTION

This paper is concerned with a problem of recognizing Textual Entailment. Textual Entailment is an important and hard basic research that can be applicable to many research fields, e.g. Question Answering, Text Summarization, Information Retrieval and Information Extraction [1]. We, LTI team, developed the following three systems where first two are the baselines: (1) Basic Element [2] based approach. (2) Voting approach combining Basic Element and more fine-grained character overlap scores. (3) Adaptable feature-based approach that combines multiple complementing features motivated by our analysis on data as well as linguistic insights. We evaluated our system in the Japanese tracks of the BC, Entrance Exam and RITE4QA subtasks at NTCIR-9 RITE [1].

In order to solve vocabulary mismatch in surface-level, we utilized large scale structured data such as WordNet and Wikipedia. The tools we built for some of sub-modules toward this goal are released as open source software for the community to use.

The rest of this paper is organized as follows. In Section 2, we will present analysis we manually conducted on the BC dev data. Section 3 and 4 will describe our baseline algorithms and proposed approach which design is motivated by the analysis in Section 2. In Section 5, tools and resources used to implement our system will be listed to help the reader's replication efforts. Section 6 will provide experimental that can hopefully give evidence that our assumptions practically works. Section 7 will report results on unseen dataset, from the NTCIR-9 RITE formal run. In Section 8, we will discuss a few observations. Finally, in Section 9, we will present concluding remarks and future works.

2. ANALYSIS OF DEVELOPMENT DATA

We analyzed the BC dev (training) dataset in order to observe general trends in the dataset and strategize possible solutions.

2.1 Manually categorizing linguistic phenomena occurrences

We analyzed hundreds of pairs and classified them into categories representing possible linguistic phenomena need to be addressed. Table 1 shows the summary of categories and the frequencies.

As expected from previous works in English community, the lexical entailment is the category with the highest frequency. Lexical entailments can be the base of sentence entailment, as you can see from an example such as the pair ID 89 in JA-BC dev dataset (see also Table 1). In the sentence ID 89, a word 自治体 */autonomous community* [1] should be entailed from 市町村 */municipality*.

However, lexical entailment alone is not sufficient to recognize sentence-level entailment. For example in the pair ID 323 (see also Table 1), all lexicons in t_2 can be entailed from lexicons in t_1, but t_2 cannot be inferred from t_1. In t_1, 沖縄/*Okinawa* modifies 基地/*base* and 米国/*U.S.* modifies 世界戦略/*world strategy*. However in t_2, 沖縄/*Okinawa* modifies 世界戦略/*global strategy* and 米国/*U.S.* modifies 基地/*base*. This dependency switch is a syntactic change and cannot be detected by the lexical-based approach. Also, negation, or polarity, is highly informative because it might change the meaning of the entire sentence.

2.2 Evidence for the need of trainable approach

We ran a simple character-overlap based entailment recognizer which is based on the following assumption: texts with high lexical overlap tend to be close in meaning, which leads to a positive entailment relationship. In Figure 1 and Figure 2, we present the histograms of the overlap ratio (the percentage of t_2 Characters contained in t_1) for each gold label. One may notice the difference in trends. In Figure 1, the Y histogram looks like a bell curve, on the other hand, N histogram isn't. There are some mass for highly overlapped N instances than it should. It may be that the data development policies [1] for JA BC N-pair affected. On the other hand, Figure 2 shows two distributions where each bell-curved Y and N distribution seem to follow the Gaussian distribution.

This analysis suggests that, even if one develops a system that works well in one dataset, there is no guarantee that it will also result in a comparable performance, even though the dataset characteristics seems to be similar.

[1] We will denote English translations in this form throughout this paper.

Table 1. Summary of manual analysis on a sample of JA BC dev dataset.

Category	Freq	Example		
		ID	t_1	t_2
Lexical Entailment	164	89	...市町村に移された/*moved to a municipality*...	...自治体に移された/*moved to a autonomous community*...
Syntactic Entailment	160	323	沖縄の基地は、米国の世界戦略と密接に結びついている。/*Military bases in Okinawa are closely related to the US global strategy*	米国の基地は、沖縄の世界戦略と密接に結びついている。/*US military bases are closely related to the Okinawa's global strategy*
Phrasal Entailment	45	25	...心をとらえてきた/*seized our heart*...	...魅了してきた/*fascinated*...
Polarity	36	390	...無駄でなかった/*not wastefull*...	...無駄だった/*wastefulness*...

Also, we can see that using the overlap score as a continuous numeric score has a certain risk. A trainable machine learning approach would not work if the feature is designed blindly. Instead, we need to go beyond a trainable approach.

Figure 1. Distribution of overlap score on JA BC dev.

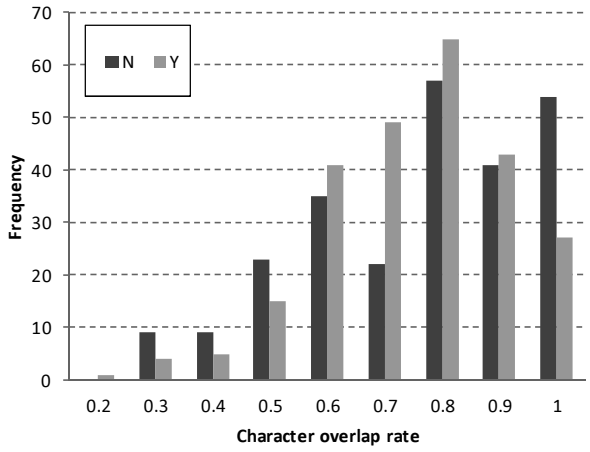
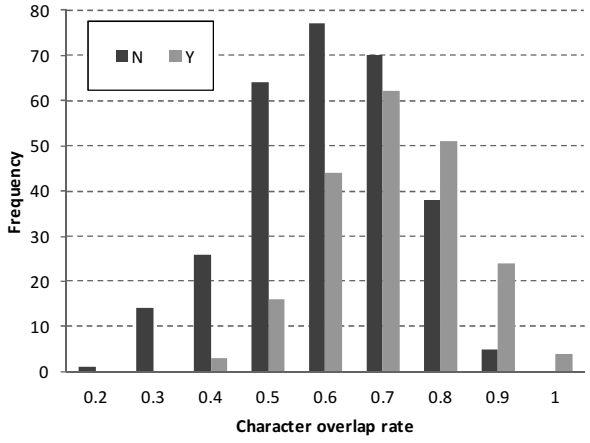


Figure 2. Distribution of overlap score on JA EXAM dev.



2.3 Toward capturing syntactic entailment

One of the difficulties in Textual Entailment recognition problem lies on the fact that it not only requires lexical entailment [3], but it also analysis on syntactic structure. As we have seen in the Table 1, a capability to recognize syntactic entailment seems to be as important as lexical entailment.

3. BASELINES

Motivated by the analysis in the previous section, we designed and implemented the components which will be described in this section in detail.

3.1 Syntactic matching with Basic Element

As a method to capture syntactic entailments, we will use Basic Element (BE) [2] which is designed to capture syntactic structure in sentence. Some studies already used it to recognize textual entailments in English texts [4].

BE is a general framework (and we will also use the term BE for an output structure from this algorithm, and a name of the feature based on BE structural matching). In this work, we re-implemented a BE implementation that operates on Japanese syntactic dependency trees [5]. In this implementation, BE extractor decomposes pieces of syntactic structures from edge and nodes in a dependency parse tree. In the rest of this subsection, we will introduce the construction of a BE, the matching criteria of two BE structures, and the sentence entailment prediction using the matching result.

3.1.1 BE construction

We used CaboCha dependency parser [6] to obtain a dependency parse tree from given a sentence. Each edge in a parse tree becomes a base of BE. A BE structure is composed of three elements, namely head, modifier and relation, which will be denoted as the following: `[head, modifier, relation]`. Head is the content word (precisely speaking, *bunsetsu* unit in Japanese) on the parent node of the edge. Modifier is the content word on the child node. Relation is the particle (often, a case marking particle, or a case marker) in the child node when the child node contains a particle (otherwise the content word in the child node fills in the relation slot). The following is an example input and output:

- **Input** (JA BC dev, ID421, t_1): 防災無線/*wireless communication system for disaster prevention* が/*GA* 停電/*power outage* で/*DE* 使えない/*out of service* 事態/*situation* が/*GA* 立ちはだかった/*stand in the way* 。
- **Output BE structures**: {[使えない/*out of service*, 防災無線/*wireless*, が/*GA*],
  [使えない/*out of service*, 停電/*power outage*, で/*DE*],
  [事態/*situation*, 使えない/*out of service*, 使えない/*out of service*],
  [立ちはだかった/*stand in the way*, 事態/*situation*, が/*GA*]}

A pseudo algorithm for BE construction is described as follows.

Algorithm 1. Basic Element construction.

```
     Input: x = parse_tree
     Output: y = be_list
 1:  be_list = empty;
 2:  for each (edge in parse_tree) {
 3:     be.modifier = get_content_word(edge.from);
 4:     be.head = get_content_word(edge.to);
 5:     if (contains_particle(edge.from)) {
 6:        be.relation = get_particle(edge.from);
 7:     } else {
 8:        be.relation = get_content_word(edge.from);
 9:     }
10:     be_list.add(be);
11:  }
12:  return be_list;
```

3.1.2 BE matching

We designed original criteria for matching BE structures, which allows soft-matching of modifiers and heads. Soft word matching is composed of three methods: Character-based matching, Kanji-based matching and Heuristic matching.

Character-based matching computes the percentage of how many characters in the t_2 word appear in the t_1 word. Examples are shown as follows:

JA BC dev, ID97.

- Input1: 雪国/*snowy county* はつらつ/*healthy and vigorous* 条例/*ordinance*
- Input2: 雪国/*snowy county* は/*WA* つらいよ/*painful* 条例/*ordinance*
- Score = 0.78

JA BC dev, ID257.

- Input1: 高性能/*high performance* 外壁材/*wall material* ダインコンクリート/*DYNE Concrete*
- Input2: 独自/*original* コンクリート/*concrete* 外壁/*wall*
- Score = 0.8

Kanji-based matching computes the ratio of Kanji in t_2 word appeared in the t_1 word. The rational is that, often, Kanji is more important than other character types, such as hiragana, katakana, numbers and English alphabets. Given the same examples as above, we obtain the score of 1.0 and 0.5 respectively.

Since the above method is not robust, a weak heuristic matching method using reading information is added to extend the Kanji-based method. The Heuristic method works as follows:

- If hypothesis word doesn't contain Kanji, return 1.
- Else if the readings match, return 1.
- Else, for each Kanji in the t_2 word, temporally remove it from t_2 and test whether Kanji in t_1 word contains Kanji in t_2 word. If any of the removals matches the words, return 1.
- Return 0 otherwise.

For each of the methods, a threshold is set to convert the overlapping ratio to binary matching decision. This threshold is set to 0.4 after tuning on the development data. The final match decision is based on the combination of the three methods.

The final BE score is calculated based on an overlap of BE structures from t_2 and t_1. The minimum and the maximum overlap ratio are learned from the development data. We submitted results from this BE baseline as one of three runs (Run ID: LTI-*-01).

3.2 Voting method

BE-based approach mainly focuses on the syntactic information. On the other hand, it is also meaningful to explore fine-grained. Voting approach is designed to combine approaches in different granularity-level, namely character-level, word-level and syntactic-level.

Character-level approach uses the same technique described in Section 2.2. Word-level approach is different from the Character-level approach in terms of tokens to be compared. We used Mecab [7] for tokenization. The Wu & Palmer semantic relatedness algorithm [8] on Japanese WordNet [9][10] is used to decide a match between words with different surface (matching criterion: relatedness score over 0.9). For both character based and word based approaches, the prediction is made by minimum and maximum overlap ratios which are learned from the development data. The third syntactic-level approach is the BE one described in the previous subsection. The voting score is given by the following formula.

$$\text{Voting-Score}(t_1,t_2) = (\text{Character-based-method}(t_1,t_2) + \text{Word-based-method}(t_1,t_2) + \text{BE-method}(t_1,t_2)\ ) / 3.$$

We submitted results from this simple voting method as one of three runs (Run ID: LTI-*-02).

4. ADAPTABLE APPROACH

In Section 2, we learned that an ideal system needs to be able to address multiple different linguistic phenomena. To this end, we decided to take a supervised machine learning approach with a careful design on features motivated by linguistics. The classification models we chose are SVM with the linear kernel [11] (in the BC subtasks) and MaxEnt [12] (in the Entrance Exam and RITE4QA subtasks). See the rational in Section 6.4.

Also, in order to deal with a dataset such as the BC subtask's where a counter-intuitive characteristics is observed, we need a certain adaptable (not only just trainable) machinery. To deal with this need, we will convert numeric continuous feature values into categorical discrete binary features.

We submitted the run result implemented with the approach described in this section as our main run (Run ID: LTI-*-03).

4.1 Features

We designed features based on two complementing principles: commonly occurring weak features and rarely occurring strong features. By strength, we mean a classification power of a feature into Y and N labels. Each feature is introduced with a rational below.

- **Morpheme Overlap** – This is a commonly occurring feature based on a morpheme overlap statistics (same as the one described in Section 2.2 where the token-level here is morpheme rather than character). Instead of exact surface-level matching, we allowed near-synonym matching described in the next subsection.
- **BE Overlap** – This is another commonly occurring feature based on overlap of BE structures described in 3.1

- **Polarity** – This feature fires when a mismatch of sentiment polarity is captured between t_1 and t_2. We assume that if one of t_1 or t_2 (but not both) has a negative modality, entailment does hold. The following negative expression cues are manually extracted for this feature from the JA BC dev dataset (they are functional words meaning "no", "not", "cannot" etc):
  ない, なし, なく, なかった, ません, できず.
- **Quote** – This feature is an N-label indicator, which fires when a quoted content in t_1 occurs in t_2. The intuition behind this feature is that what's written and what's said (or reported in quotation) have different likelihood of being true. For example, see JA BC dev ID246:
  t_1: …オジサンが「人類は麺（めん）類！」と叫んでいた。/*…the guy was shouting "mankind is a noodle-kind"*
  t_2: 人は麺（めん）類だ。/*mankind is a noodle-kind*.
  The gold label assigned to this pair by the human annotators is N. This could be one of the most difficult and may be the most controversial types in the dataset, because t_1 may *logically* entail t_2. *Pragmatically* speaking, t_2 is not a fact that can be inferable from t_1 which is apparently a joke or something. Our model is too simple to capture the differences in epistemic modalities (e.g. I think "…", I heard "…", I doubt "…") but made a contributed in this evaluation (see the next section for experimental evidence).
- **Quantification** – This feature fires when there is a mismatch in quantification expression which also indicates an N-label. The quantifier cues we extracted from the JA BC dev dataset are the following:
  限って/*only,* 限る/*only,* 限定/*only*, 唯一/*exclusively*, 必ず/*absolutely*, 常に/*always*, すべて/*all*, 全て/*all*, だけでは/*not only* 誰も/*none* ほとんど/*mostly; hardly*
- **Morpheme Diff** – The feature extractor takes a diff of sequence of morphemes from t_1 and t_2. Then, it makes an entailment recognition decision on the different morphemes using character-level heuristic soft-matching. A rational behind using this feature is that, based on an assumption that when there are only small diff between two texts, lexical entailment represents the entire text entailment. For example, see JA BC dev ID75:
  t_1: 旧経団連は、政治献金のあっせんを廃止した。/*The Former Keidanren abolished the mediation in political donations.*
  t_2: 旧経団連は、政治献金のあっせんを中止した。/*The Former Keidanren stopped the mediation in political donations.*

The diff tool detects that the different part is 廃止/*abolished* and 中止/*stopped*. Since there is a common suffix 止/*stop*, the method makes a final decision as Y, which matches the gold label.

The following table shows how many times each feature actually fired on the entire JA BC dev dataset (out of 500 pairs), categorized by the gold standard labels. We can see that both Quote and Quantification features co-occurs well the N labels (but not with Y labels).

Table 2. Feature statistics for each gold label

Feature	Y	N
Polarity	30	44
Quote	3	19
Quantification	2	11
Morpheme Diff	15	8

4.1.1 Numeric-to-binary conversion

The two overlap feature scores introduced in Section 4.1, i.e. Morpheme Overlap and BE Overlap, are non-negative real values taking a range between 0 and 1. We converted these scores to an index value in {1, …, N} which is to be used in a binary feature name.

$$f(x) = \begin{cases} 1 & \text{if } 0 \le x < \dfrac{1}{\mathrm{N}} \\ \vdots & \\ \mathrm{k} & \text{if } \dfrac{k-1}{\mathrm{N}} \le x < \dfrac{k}{\mathrm{N}} \\ \vdots & \\ \mathrm{N} & \text{if } \dfrac{N-1}{\mathrm{N}} \le x \le 1 \end{cases}$$

For instance, when the overlap score x is 0.3 and N=5, the index is 2. We experimentally found that N=5 gave us the highest accuracy on the JA BC dev dataset. When choosing the N value, one has to take the bias-variance tradeoff into consideration. When it's too low, discriminative power is small, but when it's too high, there is a risk of overfitting. See also Section 6.2 for the results.

4.2 Solving synonymy and hyponymy

Text pairs in BC subtask are created from a newswire where terminologies (e.g. abbreviations, foreign word spelling) are normalized according to a certain guideline [13]. On the other hand, we observed many alternative forms of the same lexicon between pairs in the Entrance Exam subtask, which was created from entrance exam (Daigaku Nyushi Center Shiken) and Wikipedia.

In order to solve this problem, we realized the aforementioned WordNet-based (near-)synonym resolution techniques may not be enough due to the limitation in coverage on proper nouns.

So, we used the following two additional resources created from Wikipedia which contains a lot of Named Entity entries which often lack in traditional thesauri.

- **Wikipedia Hyponymy** – This is a hypernym-hyponym resource automatically extracted from Wikipedia using NICT's Hyponymy extraction tool [14][15]. We used hypernym-hyponym data created by *hierarchy* and *category* strategies.
- **Wikipedia Redirect** – We developed a tool which can extract and utilizes Wikipedia's redirect information to be used for solving alternative forms of the same concept. For example, Wikipedia entries "Great East Japan Earthquake" and "2011 Miyagi earthquake" are both redirected to an entry "2011 Tōhoku earthquake and tsunami". In this case, we see all three as alternative forms of each other.

In the rest of this subsection, we will show the exhaustive list of term matched between t_1, t_2 (or in the other order) in JA Entrance Exam dev dataset using one of the following resources / techniques appeared in this paper: WordNet Synonymy, WordNet semantic relatedness, Wikipedia Hyponymy, and Wikipedia Redirect. Using this list, one may find how these resources may

possible help to align semantically similar or equivalent terms. We did not do any samplings; false positives are also included in natural distribution with duplicated entries removed. For the reader's convenience, we put English translations for the first 20 entries in each data.

4.2.1 WordNet synonymy

(期/*period*, 時代/*era*), (生まれる/*born*, 誕生/*birth*), (形成/*form*, 結成/*form*), (送る/*send*, 出す/*send*), (発生/, 起こる/*happen*), (商/*commerce*, 貿易/*trade*), (領域/*area*, 地域/*area*), (起きる/*happen*, 起こる/*happen*), (連合/*alliance*, 連盟/*federation*), (議長/*chairman*, 大統領/*President*), (範/*example*, モデル/*model*), (結成/formation, 組織/*organization*), (抵抗/*resist*, 反対/*opposition*), (社会/*society*, 人/*person*), (果たす/*accomplish*, 行う/*execute*), (指導/*lead*, 行う/*execute*), (装飾/*decoration*, 様式/*style*), (説話/*narrative*, 史書/*historical record*), (戦う/*fight*, 抵抗/*resist*), (合意/*agree*, 協定/*agreement*), (襲う, 起こる), (民族, 国民), (現れる, 出現), (実行, 行う), (劇, 戯曲), (構成, 組織), (連合, 同盟), (組織, 創設), (工業, 生産), (設立, 設置), (建築, 建設), (始まる, 始める), (導入, 採用), (作成, 製作), (活躍, 活動), (機関, 設立), (討つ, 破る), (文明, 文化), (裁く, 審判), (使用, 用いる), (指揮, 指導), (構成, 結成), (言葉, 語), (党, 派), (援助, 支援), (風習, 風俗), (推進, 進める), (作る, 制定), (結果, 影響), (騒動, 騒擾), (連合, 結束), (制限, 抑制), (承認, 調印), (設置, 発足), (署名, 調印), (講和, 平和), (出席, 参加), (規定, 明記), (廃止, 撤廃), (広まる, 展開), (会, 組合), (組合, 同盟), (確立, 結成), (護衛, 防衛), (設置, 設ける), (編制, 編成), (法, 規則), (制定, 設置), (祭祀, 祭り), (使う, 用いる), (営む, 行う), (戦う, 交戦), (記録, 記す), (考える, 見る), (徴税, 租税), (時代, 期), (規定, 法), (揉め事, 紛争), (職務, 任務), (制定, 定める), (建設, 造営), (多数, 多く), (所, 地), (際, 時), (節減, 節約), (選出, 選ぶ), (判断, 審査), (守る, 保護), (許す, 認める), (裁判所, 裁判官), (裁判官, 裁判所), (指名, 任命), (救う, 救済), (事件, 場合), (被害, 毀損), (指定, 明示), (原則, 法), (増加, 増える), (成長, 振興), (力, 補助), (統合, 合併), (経営, 行う), (料, 負担), (負担, 料), (保険, 保障), (反する, 違反), (自己, 自ら), (条例, 法律), (定める, 制定), (確立, 制定), (別, 相違), (模倣, 複製), (児童, 子ども), (階級, 等), (選ぶ, 選択), (事業, 企業), (取引, 条約), (クジラ, 鯨), (論争, 対立), (保存, 保護), (位置付ける, 位置づける), (分別, 分ける), (分類, 分ける), (緊縮, 経済), (推進, 促進), (集団, 部門), (憲章, 憲法), (合致, 承認), (明記, 規定), (法, 規定), (名づける, 呼ぶ), (行動, 態度), (相違, 違い), (直面, 向き合う), (開発, 発展), (国民, 民族), (結成, 設立), (削減, 制限), (限定, 制限), (受ける, 行う), (返還, 返済), (発表, 刊行), (発表, 出版), (禁止, 禁じる), (選挙, 選出), (用いる, 利用), (受ける, 認める), (行う, 実施), (賛成, 可決), (可決, 議決), (採決, 議決), (持つ, 有する), (選出, 選挙), (団体, 組織), (組織, 団体), (賛成, 同意), (先, 所), (雇用, 労働), (勤労, 労働), (通貨, 金), (経済, 融資), (支援, 援助), (維持, 保護), (設備, 施設), (制度, 仕組み), (制定, 設ける), (削減, 減少), (投資, 経済), (経済, 投資), (成長, 発展), (移す, 転換), (場合, 際), (損害, 害する)

4.2.2 WordNet semantic relatedness

(誕生/*birth*, 生まれる/*born*), (ポリス/*police*, 軍/*military*), (ベール/*veil*, 保護/*protection*), (孫/grand-son, 帝/emperor), (行う/execute, 果たす/*accomplish*), (流通/*circulation*, 流入/*inflow*), (生産/*production*, 構成/*composition*), (帝/*emperor*, 始祖/*earliest ancestor*), (打ち壊す/*shatter*, 破壊/*destroy*), (前半/*first-half*, 後半/*last-half*), (聖/*saint*, 皇帝/*emperor*), (末/*end*, 死/*death*), (大帝/*the Great*, 人/*person*), (国王/*king*, 皇帝/*emperor*), (進出/*foray*, 移動/*movement*), (措置/*measure*, 対策/*counter measure*), (総/*gross*, 13/*13*), (展開/*expansion*, 広まる/*spread*), (交戦/*war*, 戦う/*fight*), (初期/*initialization*, 設置/*installation*), (所領, 権限), (禁止, 法), (定める, 制定), (救済, 救う), (投票, 選挙), (選挙, 投票), (憲法, 法律), (法律, 事項), (都市, 環境), (首長, 市長), (違反, 反する), (制定, 定める), (議定, 締結), (全般, 町), (規定, 概念), (向き合う, 直面), (禁止, 条約), (抑制, 封じ込め), (貿易, 輸出), (輸出, 貿易), (原油, 石油), (返済, 返還), (禁じる, 禁止), (情報, データベース), (国, 町), (有する, 持つ), (憲法, 条約), (議員, 立候補者), (法律, 条項)

4.2.3 Wikipedia hyponymy

(共和国/*republic*, フィリピン/*Philippines*), (冤罪/ *false accusation*, ドレフュス事件/*Dreyfus affair*), (ロシア連邦大統領/*President of Russia*, ロシアの大統領/*President of Russia*), (軍事/*military*, 軍隊/*armed forces*), (僧/*sangha*, 上座/*kamiza*), (徳川家康/*Tokugawa Ieyasu*, 徳川氏/*Tokugawa clan*), (軍事/*military*, 治安/*public safety*), (貧困/*poverty*, 飢饉/*famine*), (農業/*agriculture*, 農地/*arable land*), (福祉/*welfare*, 介護/*elderly care*), (地方公共団体/*local government*, 地方自治/*local self-government*), (温室効果ガス/*greenhouse gas*, 二酸化炭素/*carbon dioxide*), (空売り/*short selling*, 金融/*finance*), (軍事/*military*, 社会/*society*), (人権/*human rights*, 自由/*freedom*), (人権/*human rights*, 権利/*rights*), (データ/*data*, 情報/*information*), (地方/*region*, 地域/*area*)

4.2.4 Wikipedia Redirect

(アナトリア半島/*Anatolia*, 小アジア/*Asianic*), (ギリシャ/*Greece*, ギリシア/*Greece*), (1/*1*, 第一/*first*), (平和条約/*peace treaty*, 講和/*pacification*), (ヴァスコ・ダ・ガマ/*Vasco da Gama*, ヴァスコ=ダ=ガマ/*Vasco da Gama*), (イスラム/*Islam*, イスラーム/*Islam*), (イスラム教/*Islam*, イスラーム教/*Islam*), (マリ帝国/*Mali Empire*, マリ王国/*Mali Empire*), (アメリカ人/*American*, アメリカ合衆国/*U.S.A.*), (アパルトヘイト/*apartheid*, 人種隔離政策/*racial segregation policy*), (五カ年計画/*5-year-plan*, 五か年計画/*5-year-plan*), (2 月/*February*, 二月/February), (ソビエト/*Soviet*, ソヴィエト/*Soviet*), (ソビエト連邦/*Soviet Union*, ソヴィエト連邦/*Soviet Union*), (ロシア連邦大統領/*Russian President*, ロシアの大統領/*President of Russia*), (マフディー戦争/*Mahdist War*, マフディー運動

/*Mahdist War*), (青年トルコ人革命/*Young Turk Revolution*, 青年トルコ革命/*Young Turk Revolution*), (宣教/*missionary* , 布教/propagandism), (汎ゲルマン主義/*Pan-Germanism*, パン＝ゲルマン主義/*Pan-Germanism*), (ラッダイト運動/*Luddite movement*, ラダイト運動/*Luddite movement*), (破壊, 壊す), (アユタヤ王朝, アユタヤ朝), (中華ソビエト共和国, 中華ソヴィエト共和国), (イスラム王朝, イスラーム王朝), (イルハン朝, イル＝ハン国), (マジャル人, マジャール人), (ハールーン・アッ＝ラシード, ハールーン＝アッラシード), (龍樹, ナーガールジュナ), (龍樹, 竜樹), (龍, 竜), (サーサーン朝, ササン朝), (受け, うけ), (甲申政変, 甲申事変), (澤, 沢), (福澤諭吉, 福沢諭吉), (第二次大戦, 第二次世界大戦), (安保条約, 安全保障条約), (子供, 子ども), (人間, 人々), (律令制, 律令国家), (徳川家, 徳川氏), (天領, 幕領), (大名屋敷, 武家屋敷), (天保の大飢饉, 天保の飢饉), (在外選挙, 在外投票), (POS システム, 販売時点情報管理), (ショップ, 店), (ブック, 本), (市長, 市町村長), (地方公共団体, 地方自治体), (可決, 決議), (地方公共団体の長, 首長), (三位一体の改革, 三位一体改革), (知的財産権, 知的所有権), (争い, 紛争), (社会企業家, 社会起業家), (動植物, 生物), (自然, 野生), (国際連合, 国連), (国際連合安全保障理事会, 国連安全保障理事会), (子供, 子ども), (開発途上国, 発展途上国), (国際連合安全保障理事会, 安保理), (ソビエト連邦, ソ連), (国際連合貿易開発会議, 国連貿易開発会議), (国際連合人間環境会議, 国連人間環境会議), (国際連合環境計画, 国連環境計画), (憲法改正, 憲法の改正), (年棒, 年俸), (施策, 政策), (農家, 農業従事者), (米, コメ), (特定商取引に関する法律, 特定商取引法)

5. TOOLS AND RESOURCES USED

The approaches described in the previous section are implemented using the set of software our team has implemented and released: JAWJAW, WS4J and Wikipedia Redirect. We also used some existing tools and resources that are publicly available. A summary of tools and resources used in our system is described in the following table.

Table 3. Tools and resources used.

Tool/Resource	Description
CaboCha [6][2] on MeCab [7][3]	Syntactic dependency parser for Japanese, which internally calls a morphological analysis tool (in our case MeCab).
Hyponymy extraction tool [14][15]	We created a dictionary of hypernym-hyponym pairs from Japanese Wikipedia using this tool.
JAWJAW[4] on Japanese WordNet [10][5]	We used this tool to find synonyms of a word.
WS4J[6]	WordNet Similarity implementation for Java, which includes a metric by Wu and Palmer [8].
Wikipedia Redirect[7]	This tool can generate a dictionary of page title and its redirection from Wikipedia.
MinorThird[8] on libsvm[9] and iitb.CRF[10]	MinorThird provides an interface to various kinds of machine learning algorithms. We used SVM [11] and MaxEnt [12] implementations wrapped in MinorThird.

6. EXPERIMENTS

We conducted in-house pre-formal-run experiments on dev data in order to measure the contribution of each feature used in the adaptable textual entailment system proposed in the previous section.

6.1 Ablation study

Table 4 shows all-but-one ablation results where one feature is removed at a time. We did 5-fold cross validation experiment on JA BC and EXAM dev dataset using a SVM (Linear Kernel) classifier. The absolute difference from the all-feature settings is interpreted as the following: larger negative number played more important role. All the features achieved non-positive diff numbers, which supports the usefulness of the features.

Table 4. Experiment result: all-but-one feature ablation.

Feature	BC		EXAM	
	Acc	Diff	Acc	Diff
All features	*62.6%*	*N/A*	*68.9%*	*N/A*
- Morpheme Overlap	61.0%	-1.6%	59.1%	-9.8%
- BE Overlap	54.2%	-8.4%	68.9%	0.0%
- Quote	61.4%	-1.2%	68.7%	-0.2%
- Polarity	59.8%	-2.8%	68.7%	-0.2%
- Quantification	62.2%	-0.4%	68.9%	0.0%
- Morpheme Diff	57.2%	-5.4%	68.7%	-0.2%

Notice that the features impacted the most is BE Overlap and Morpheme Overlap in BC and EXAM, respectively. Also, Morpheme Overlap is much more useful than any other features in EXAM, whereas this trend does not exist in the BC results. This difference may imply that it is advantageous to use an adaptable technique for recognizing textual entailment.

6.2 Numeric-to-binary conversion

We verified if numeric-to-binary conversion of features contributed to our system, in 5-fold cross validation using a SVM with the linear kernel. We observed a large difference between runs with or without conversion on BC (see Table 5). It suggests that, to handle bumpy distributions (such as BC's seen in Figure 1), numeric-to-binary conversion is effective.

[2] http://code.google.com/p/cabocha/
[3] http://code.google.com/p/mecab/
[4] http://code.google.com/p/jawjaw/
[5] http://nlpwww.nict.go.jp/wn-ja/index.en.html
[6] http://code.google.com/p/ws4j/
[7] http://code.google.com/p/wikipedia-redirect/
[8] http://minorthird.sourceforge.net/
[9] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[10] http://crf.sourceforge.net/

Table 5. With or without numeric-to-binary conversion.

Run	Acc	
	BC	EXAM
All features *with* conversion	62.6%	68.9%
All features *without* conversion	56.0%	69.9%

6.3 Adaptability

We trained the system on JA BC dev data and evaluated on JA EXAM dev dataset. The accuracy was 51.7%, which is much less than expected. We spent several minutes training the system on EXAM dev dataset, and obtained 69.5%. This result suggests there is no guarantee that a system performed good in one datset performs similarly well in another dataset. At the same time, we learned our system can quickly adapt to a different dataset.

Table 6. Experiment result: Adaptability experiment.

Dataset Trained	Dataset Tested	Acc
BC	BC	*62.6%
BC	EXAM	51.7%
EXAM	EXAM	*69.5%

* Evaluated in 5-fold cross validation.

6.4 Comparison of machine learning algorithms

In Table 7, we compared multiple major machine learning classification models on JA BC/EXAM dev dataset using all features. In the BC Entrance Exam subtask, we used SVM with the linear kernel as it is simple (as compared to using other kernels, which has advantage in training/runtime speed as well) and effective. In the Entrance Exam and RITE4QA subtasks where we needed to generate a confidence scores in addition to labels, we used MaxEnt because it is a probabilistic model that generates a probability between 0 and 1 in a natural way[11].

Table 7. Experiment result: comparison of machine learning models.

Classifier	Acc	
	BC	EXAM
Decision Tree	57.6%	66.9%
Margin Perceptron	59.4%	68.5%
MaxEnt	58.8%	**69.5%**
Naïve Bayes	57.0%	68.7%
SVM (Linear Kernel)	**62.6%**	68.9%
SVM (RBF Kernel)	62.6%	68.7%
SVM (Polynomial Kernel)	61.4%	68.9%
SVM (Sigmoid Kernel)	62.6%	68.9%
Voted Perceptron	59.6%	70.1%

7. FORMAL RUN RESULT

The following are the runs we submitted to NTCIR-9 RITE.

- Run 01: BE baseline
- Run 02: Voting system baseline
- Run 03: Adaptable approach

The formal run results, together with results on dev datasets, are shown in Table 8 where our adaptable approach (bold face) constantly outperforms the two strong baselines.

[11] It is also possible to generate an estimation of probability using margin-based classification models [16] though.

Table 8. Formal run result.

Run	Dev	Test (formal run)		
	Acc	Acc	Top1	MRR
BC-01	57.6%	53.4%	-	-
BC-02	61.2%	54.2%	-	-
BC-03	***62.6%**	**54.6%**	-	-
EXAM-01	61.1%	60.2%	-	-
EXAM-02	67.5%	65.4%	-	-
EXAM-03	***69.5%**	**66.7%**	-	-
RITE4QA-01	-	****84.3%**	12.7%	22.2%
RITE4QA-02	-	****64.1%**	17.4%	25.6%
RITE4QA-03	-	****67.5%**	**21.4%**	**29.8%**

* Evaluated in 5-fold cross validation.
** Accuracy is the secondary metric in the RITE4QA subtask

8. DISCUSSION

8.1 Overfitting

In the BC subtask, dev and test data were created from the same pool and randomly split into two [1]. There is overfitting, which is a phenomenon where test performance is worse than training performance, in our system as seen in the formal run results in Table 8. This suggests that we need to elaborate more on the generality of feature extractor implementation (e.g. with more cues), although the feature design look already general.

8.2 A case study: when naïve approach with WordNet might fail

In an ordinary context of upward monotonicity, a concept can be expanded to its superset without losing the statement's validity. For example, "日本で地震が起きた/*There was an earthquake in Japan*" would entail "アジア（のある国）で地震が起きた/*There was an earthquake in (one of countries in) Asia*", since "アジア/*Asia*" is a hypernym of "日本/*Japan*". On the contrary, "アジアで地震が起きた/*There was an earthquake in Asia*" would not entail "日本で地震が起きた/*There was an earthquake in Japan*". Thus, a naive approach to deal with this type of sentence pairs is to check if the text and the hypothesis contain corresponding hyponym and hypernym, respectively. However, there are situations where the sentence pairs can contain irrelevant hyponym-hypernym pairs. Consider JA Entrance Exam dev ID31:

t_1: イギリス国教会は、16 世紀のイングランド王ヘンリー8 世からエリザベス 1 世の時代にかけてローマ教皇庁から離れ、独立した教会となったものである。/*The Church of England is a church that became independent from the Pope in the era of the King of England Henry VIII through the Queen Elizabeth I in the 16th century.*

t_2: 16 世紀にイギリスでは、ヴィクトリア女王によって、イギリス国教会が確立された。/*In England, the Church of England has been established by the Queen Victoria.*

Although t_1 does not entail t_2, WordNet would detect 時代/*age; era* and 世紀/*century* to be in a hypernym-hyponym relationship if applied blindly. Therefore, we expect an approach that extract hyponym-hypernym relations based on aligned words would outperform a naïve approach with a bag-of-words representation.

9. CONCLUSION AND FUTURE WORKS

We presented the LTI's system participated in NTCIR-9 RITE. Through analysis, we assumed that multiple linguistic phenomena must be captured, and there is a need of adaptability in a Textual Entailment recognition system. We experimentally showed that they are reasonable assumptions to make. Our contribution also includes releasing open source software WS4J and Wikipedia Redirect.

Our future works include detailed analysis with more detailed categorizations, such as the ones seen in [17] that classified kinds of common knowledge needed for recognizing Textual Entailment. Another future work is to elaborate more on capturing linguistic modalities. Especially, recognizing epistemic modality, or committed and non-committed belief [18] could be a sophisticated extension of Quote feature we used.

10. REFERENCES

[1] Shima, Hideki, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi, and Koichi Takeda. 2011. Overview of NTCIR-9 RITE: Recognizing Inference in TExt. to appear, In Proceedings of NTCIR-9 Workshop, Japan.

[2] Hovy, Eduard, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC)

[3] Glickman, Oren, Ido Dagan, and Moshe Koppel. 2005. Web based probabilistic textual entailment. In Proceedings of the 1st RTE Workshop, Southampton, UK, 2005.

[4] Nicholson, Jeremy, Nicola Stokes, and Timonthy Baldwin. 2006. Detecting Entailment Using an Extended Implementation of the Basic Elements Overlap Metrics. In Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, pages 122-127, Venice, Italy, April 2006.

[5] Fukumoto, Junichi. 2007. Question answering system for non-factoid type questions and automatic evaluation based on BE method, Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access (NTCIR-6), Tokyo, Japan, 2007, pp. 441-447.

[6] Kudo, Taku and Yuji Matsumoto. 2002. Japanese Dependency Analysis using Cascaded Chunking. In Proceeding of the 6th Conference on Natural Language Learning, pages 63-69, 2002.

[7] Kudo, Taku, Kaoru Yamamoto, Yuji Matsumoto. 2004. Appliying Conditional Random Fields to Japanese Morphological Analysis. In Proceedings of EMNLP 2004.

[8] Wu, Zhibiao and Martha Palmer. 1994. Verb semantics and lexical selection. In 32nd. Annual Meeting of the Association for Computational Linguistics, pages 133–138, New Mexico State University, Las Cruces, New Mexico

[9] Miller, George A. 1995. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[10] Isahara, Hitoshi, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama and Kyoko Kanzaki. 2008. Development of Japanese WordNet. In Proceedings of LREC-2008, Marrakech.

[11] Vapnik, Vladimir. N. 2000. The nature of statistical learning theory. Springer.

[12] A. Berger, V. Della Pietra, and S. Della Pietra. 1996. A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39-71, 1996.

[13] The Mainichi Newspapers Co.,Ltd. (Eds.) 2007. Mainichi Shimbun Yougo Shu. ISBN-13: 978-4620317953. (In Japanese)

[14] Sumida, Asuka, Naoki Yoshinaga and Kentaro Torisawa. 2008. Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia. In Proceedings of the Sixth International Language Resources and Evaluation, 2008.

[15] Oh, Jong-Hoon, Kiyotaka Uchimoto and Kentaro Torisawa. 2009. Bilingual Co-Training for Monolingual Hyponymy-Relation Acquisition. In Proceedings of ACL-IJCNLP-2009, pp.432-440.

[16] Wu, Ting-Fan, Chin-Jen Lin, and Ruby C. Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research, 5:975-1005, 2004.

[17] LoBue, Peter and Alexander Yates. 2011. Types of Common-Sense Knowledge Needed for Recognizing Textual Entailment. In Proceedings of ACL 2011.

[18] Diab, Mona, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, Weiwei Guo. 2009. Committed Belief Annotation and Tagging. In Proceedings of ACL-IJCNLP 2009 Workshop on The Third Linguistic Annotation Workshop (The LAW III).