International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.

易繫辭曰上古結繩而
治後世聖人易之以書
契百官以治萬民以察
說文敘曰蓋文字者經
藝之本宣教明化之始
前人所以垂後後人所
以識古故曰本立而道
生知天下之至賾而不
可亂也教化既萌文心
雕龍則謂人之立言因
字而生句積句而成章
積章而成篇篇之彪炳

Vol.19 No.3 September 2014 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 19, no. 3, September 2014

Contents

Papers

The Association for Computational Linguistics and Chinese Language Processing

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 3, September 2014, pp. 1-24

BCCWJ-TimeBank: Temporal and Event Information Annotation on Japanese Text

Masayuki ASAHARA[*], **Sachi KATO**[*], **Hikari KONISHI**[*],

Mizuho IMADA[*], **and Kikuo MAEKAWA**[*]

Abstract

Temporal information extraction can be divided into the following tasks: temporal expression extraction, time normalisation, and temporal ordering relation resolution. The first task is a subtask of a named entity and numeral expression extraction. The second task is often performed by rewriting systems. The third task consists of event anchoring. This paper proposes a Japanese temporal ordering annotation scheme that is used to annotate expressions by referring to 'the 'Balanced Corpus of Contemporary Written Japanese' (BCCWJ). We extracted verbal and adjective event expressions as <EVENT> in a subset of BCCWJ and annotated a temporal ordering relation <TLINK> on the pairs of these event expressions and time expressions obtained from a previous study. The recognition of temporal ordering by language recipients tends to disagree with the normalisation of time expressions. Nevertheless, we should not strive for unique gold annotation data in such a situation. Rather, we should evaluate the degree of inter-annotator discrepancies among subjects in an experiment. This study analysed inter-annotator discrepancies across three annotators performing temporal ordering annotation. The results show that the annotators exhibit little agreement for time segment boundaries, whereas a high level of agreement is exhibited for the annotation of temporal relative ordering tendencies.

Keywords: Temporal Information Processing, Event Semantics, Corpus Annotation.

[*] National Institute for Japanese Language and Linguistics, Japan
E-mail: masayu-a@ninjal.ac.jp

1. Introduction

Temporal information processing in natural language texts has received increasing scholarly attention in recent years. Since the temporal orders of events often have implications for causal relations (cause and effect), identifying them is an essential task for deep understanding of language. Several types of resources for English temporal information processing have been developed, including an annotation specification TimeML (Pustejovsky *et al.*, 2003a) and annotated corpora, such as TimeBank (Pustejovsky *et al.*, 2003b) and Aquaint TimeML Corpus.

The English annotation specification has been extended as an International Standard Organization (ISO) standard of a temporal information mark-up language, namely ISO TimeML (Pustejovsky, Lee, Bunt, & Romary, 2010), which covers Italian, Spanish, Chinese, and other languages. Temporal information-annotated corpora in various languages have been developed and have been shared by natural language processing researchers. TempEval-2 (Verhagen, Sauri, Caselli, & Pustejovsky, 2010), a task for the SemEval-2010, and TempEval-3 (UzZaman *et al.*, 2013), a task for the SemEval-2013, have been proposed as shared temporal-relation reasoning tasks. In these shared tasks, datasets for English, Italian, Spanish, Chinese, and Korean are provided.

Nevertheless, there is no such resource for the Japanese language. In this paper, we present a means of porting a subset of ISO-TimeML into the Japanese language and describe the basic specifications of ‘BCCWJ-TimeBank,’ which is a realisation of the temporal information annotation of the ‘Balanced Corpus of Contemporary Written Japanese,’ or BCCWJ (Maekawa, 2008).

The purposes of temporal information annotation differ in accordance with the research goal. For natural language processing, users may need to develop and evaluate analysers for the annotation. Hence, for linguistic purposes, some users may want to formalise semantic expressions for temporal and event information expressions. Other users may want to evaluate human cognitive processes related to the expressions. The former purpose requires unique and consistent annotations. The latter purpose does not require consistent annotation; instead, it may require ways to evaluate variation among annotations. In this study, we conduct ‘pair annotation’ to formalise semantic expressions for temporal expressions. We also perform evaluation of the cognitive process of the expressions for temporal ordering annotation.

Porting TimeML into other languages can be challenging because of differences between languages. Even if we made a standardisation of ISO-TimeML, there would still be slight differences among the resources in terms of annotation targets, styles, formalism, philosophy, objectives, and focuses. Our research target is ‘temporal ordering’ in Japanese documents. We want to establish a machine learning-based temporal ordering analyser of the event and time expressions in Japanese. Before we develop the temporal ordering analyser, however, we have to analyse how well human annotators detect the temporal ordering. Therefore, our main research question in this article is to evaluate human annotators' skills in Japanese temporal ordering annotation. We permit inconsistency in semantic-level annotation among the annotators and quantitatively evaluate this inconsistency for the type of temporal ordering. Under the main research question, we did not perform sound and complete localization of ISO-TimeML to Japanese, as previous research has done.

The contributions of this work are as follows. First, to the best of our knowledge, this is the first corpus-based study on Japanese temporal information annotation. Second, we introduce two annotation paradigms for linguistic research. One paradigm is 'pair programming'-like annotation for consistent annotation. The other paradigm is annotation as a subjective experiment. Third, we evaluate cognitive processes in human temporal information processing.

The rest of this paper is organised as follows. Section 2 discusses previous studies related to this work. Section 3 briefly presents our annotation specification. Section 4 outlines the annotation processes of our work. Section 5 presents the corpus statistics and evaluations. Finally, Section 6 concludes this paper.

2. Previous Studies

This section discusses previous studies on BCCWJ-TimeBank. Section 2.1 presents English temporal information processing. Section 2.2 presents ISO-TimeML, which is a standardisation of the annotation schema. Section 2.3 presents Asian language resources related to temporal information processing. Section 2.4 explains the target corpus of Japanese.

2.1 English Temporal Information Processing

MUC-6 (Grishman & Sundheim, 1996) was a workshop on information extraction, which included temporal expression extraction as a subtask. TimeML <TIMEX3> tags used to be the *de facto* standard of normalization of temporal expressions; however, temporal information processing was then extended to event semantics. TimeML provides an annotation schema for event expressions and temporal relation extraction. Following this, TimeBank and some other corpora were developed. Using this corpus, machine learning-based temporal relation extraction methods have been developed (Boguraev & Ando, 2005; Mani, 2006). In addition, shared task workshops, including TempEval (Verhagen *et al.*, 2007), TempEval-2 (Verhagen *et al.*, 2010), and TempEval-3 (UzZaman *et al.*, 2013), have been held in more formalized evaluation settings.

2.2 ISO-TimeML: Standardisation of the Annotation Schema

The ISO Technical Committee (TC 37) has proposed several standards for language resources under the collective category ‘Terminology and other language and content resources’. The committee (SC) is divided into four areas. TC 37/SC 4 is charged with looking at annotation standards for all areas of natural language resources. This area includes six working groups (WG) to design language annotation specification mark-up languages, such as stand-off mark-up and XML. TC 37/SC 4/WG 2, the semantic annotation WG, discusses semantic annotation standards. The original TimeML developers and TC 37/SC 4/WG 2 defined ISO-TimeML as a Semantic Annotation Framework (SemAF)-Time, that is ISO-24617-1:2012, within the context of TC 37/SC 4.

TimeML and ISO-TimeML define four types of entities: <TIMEX3>, <EVENT>, <MAKEINSTANCE>, and <SIGNAL>. The <TIMEX3> tag specifies various attributes of time expressions, such as tid, type, quant, freq, mod, and value. The time expressions are categorized into four types: DATE, TIME, DURATION, and SET. The attribute @value includes the normalised values of the time expressions in a machine-readable format. The <EVENT> tag specifies various attributes of event expressions, including the class of the event, tense, grammatical aspect, polarity, and modal information. The <MAKEINSTANCE> tag presents the event instances expressed by <EVENT>-tagged expressions. Finally, the <SIGNAL> tag annotates elements to indicate how temporal objects are related amongst themselves.

TimeML and ISO-TimeML also define several types of links. Among these, <TLINK> expresses temporal order among instances of time expressions, event expressions, or both.

2.3 Time Information Annotation in Asian Languages

Japanese temporal information processing is still being developed. We have developed temporal expression extraction resources only as a subtask of named entity extraction. The IREX NE Task (Sekine & Isahara, 2000) includes time expressions as the target. Sekine, Sudo, and Nobata (2002) maintained an extended named entity hierarchy for Japanese and other languages, which includes five subcategories of time expressions and six subcategories of period expressions.

In the case of Chinese, Cheng developed a Time-ML compatible Chinese Temporal Annotation Corpus (Cheng *et al*., 2008a) and proposed some models for the data (Cheng *et al*., 2008b). This was the first localization work with TimeML for Chinese and was before ISO-TimeML. Cheng performed temporal ordering information annotation on 151 articles from the Penn Chinese Treebank (Xue *et al*., 2005). In their work, syntactic dependency relations derived from the Penn Chinese Treebank were utilized for the annotation and temporal ordering estimation. There are two representative temporal information annotated corpora in addition to the abovementioned works: the Chinese part of the ACE 2005 multilingual training corpus (Walker *et al*., 2006) and the TempEval-2 Chinese data sets (Xue & Zhou, 2010). The former is only for temporal expression extraction and normalization. The latter focuses on the temporal ordering of four relations, similar to this work, and it will be presented in Section 3.4. The TempEval-2 Chinese data sets are also based on 60 articles from the Penn Chinese Treebank that were analysed by a two-phase double blind and adjudication process. Nevertheless, sound and complete annotation cannot be achieved. Recently, the temporal expression annotation of TempEval-2 Chinese data was fixed by Li *et al*. (2014), and a Chinese temporal tagger based on the new annotation is publicly available.

In the case of Korean, KTimeML is a temporal information annotation guideline for Korean (Im *et al*., 2009). Im and his colleagues utilized morpheme-based stand-off annotation and surface-based annotation.

A contrastive evaluation among Asian temporal information language resources is difficult. The research objectives vary among the previous articles. In addition, some detailed annotation guidelines are not in the previous articles but are written in manuals in their own language. Nevertheless, we emphasize that the temporal information processing of Asian languages is still an ongoing process. There are no sound and complete language resources on temporal information processing. One reason might be that the localization of all ISO-TimeML tags and attributes does not always help the temporal information processing. In addition, there are still language-independent issues for each language. Another reason might be that the human recognition system of temporal information is not stable among people. This article attempts to evaluate the human recognition system of temporal ordering in cognitive experiments.

2.4 BCCWJ and its Annotations

BCCWJ was publicly released in 2011 by the National Institute for Japanese Language and Linguistics (NINJAL) in Japan. It consists of three sub-corpora: ‘Publication,’ ‘Library,’ and ‘Special purpose’. ‘Publication’ consists of samples extracted randomly from books, magazines, and newspapers published during 2001-2005. ‘Library’ consists of randomly extracted samples from texts in circulation at libraries during the period 1986-2005. Finally, the ‘Special purpose’ sub-corpus consists of several mini-corpora without a statistical sampling method being used. It includes text from Yahoo! Answers, Yahoo! Blogs, white papers, and school textbooks. The total size of BCCWJ is about 100 million words.

The part of BCCWJ called ‘CORE’ manually annotates word boundaries, base phrase boundaries, and morphological information. CORE consists of six registers found in ‘Publication’ and ‘Special purpose’: books (PB), magazines (PM), and newspapers (PN) from ’Publication,’ along with Yahoo! Answers (OC), Yahoo! Blogs (OY), and white papers

(OW) from ‘Special purpose’. The size of CORE is about 1.3 million words.

The BCCWJ data include several annotations, such as metadata, document structure, sentence boundaries, word boundaries, and phrase (bunsetsu) boundaries. NINJAL suggests two sorts of word delimitation definitions: one is a *short word unit* and the other is a *long word unit*. Each of these word delimitations is coded with the UniDic part-of-speech tag (Den *et al.*, 2008).

Several research institutes have developed further linguistic annotations for CORE, such as syntactic dependency structures, developed by Nara Institute of Science and Technology (NAIST) and NINJAL; predicate-argument relations, developed by NAIST; named entities, developed by Tokyo Institute of Technology (TITECH); modality, developed by Tohoku and Yamanashi Universities; and Japanese framenet, developed by Keio University. The multi-word functional expressions are maintained by Tsukuba University. The CORE samples are split into annotation priority sets from A to E to allow the annotations to overlap as much as possible. Some of these annotations can be used as presupposed information for our annotation. In this way, BCCWJ is the most promising resource making linguistic annotations for both NLP and linguistic researchers.

Table 1 shows the basic statistics and priority sets of BCCWJ CORE. The word unit is based on the *short word unit*, UniDic standard (Den, Nakamura, Ogiso, & Ogura, 2008); UniDic is a lexicon for Japanese morphological analysis.

Table 1. BCCWJ CORE: Registers and priority sets.

Register	(Abbr.)	Priority Set	# of Samples	# of Words
White Paper	OW	A to D	62	197,011
Books	PB	A to D	83	204,050
Newspapers	PN	A to E	340	308,504
Yahoo! Answers	OC	A to B	938	93,932
Magazines	PM	A to D	86	202,268
Yahoo! Blog	OY	A to B	471	92,746

3. Temporal Information Annotation Specification

This section presents a specification for Japanese temporal information annotation. The annotation is realised as BCCWJ-TimeBank. The specification is based on TimeML (Pustejovsky *et al.*, 2003a) and is adapted for the Japanese language. Figure 1 shows an example of the annotation. Below, we present an overview of the specification of TimeML tags: <TIMEX3> for temporal expressions, <EVENT> and <MAKEINSTANCE> for event expressions, and <TLINK> for temporal ordering. We mention other tags that we exclude

PN23_00001 Sample in BCCWJ CORE

```
<TIMEX3 @value="2002-04-11" @definite="true" @tid="t0"  functionInDocument="CREATION_TIME"
type="DATE"/><sentence>地方自治体が<EVENT @class="NULL" @eid="e25">運営する</EVENT>公営地下鉄二十六
路線のうち<TIMEX3 @value="FY2000" @definite="FALSE" @valueFromSurface="FY2000" @tid="t4"
@type="DATE">二〇〇〇年度</TIMEX3>決算で経常損益が黒字なのは、札幌市南北線など四路線に<EVENT
@class="I_ACTION" @eid="e26">とどまった</EVENT>ことが、公営交通事業協会が<TIMEX3 @value="2002-04-10"
@definite="true" @valueFromSurface="XXXX-XX-10" @tid="t5"
type="DATE">十日</TIMEX3><EVENT @class="I_ACTION" @eid="e27">まとめた</EVENT>報告書で<EVENT
@class="I_STATE" @eid="e28">分かった</EVENT>。</sentence>
<MAKEINSTANCE @eventID="e26" @eiid="ei26"/>
<MAKEINSTANCE @eventID="e27" @eiid="ei27"/>
<MAKEINSTANCE @eventID="e28" @eiid="ei28"/>
<TLINK @relTypeA="after" @relTypeB="after" @relTypeC="during" @task="DCT"
@timeID="t0" relatedToEventInstance="ei26"/>
<TLINK @relTypeA="after" @relTypeB="after" @relTypeC="after" @task="DCT"
@timeID="t0" @relatedToEventInstance="ei27"/>
<TLINK @relTypeA="after" @relTypeB="after" @relTypeC="after" @task="DCT"
@timeID="t0" @relatedToEventInstance="ei28"/>
<TLINK @relTypeA="vague" @relTypeB="equal" @relTypeC="during" @task="T2E"
@timeID="t4" @relatedToEventInstance="ei26"/>
<TLINK @relTypeA="vague" @relTypeB="before" @relTypeC="before" @task="T2E"
@timeID="t4" @relatedToEventInstance="ei27"/>
<TLINK @relTypeA="vague" @relTypeB="before" @relTypeC="before" @task="T2E"
@timeID="t4" @relatedToEventInstance="ei28"/>
<TLINK @relTypeA="after" @relTypeB="before" @relTypeC="during" @task="T2E"
@timeID="t5" @relatedToEventInstance="ei26"/>
<TLINK @relTypeA="contains" @relTypeB="after" @relTypeC="finishes" @task="T2E"
@timeID="t5" @relatedToEventInstance="ei27"/>
<TLINK @relTypeA="contains" @relTypeB="equal" @relTypeC="before" @task="T2E"
@timeID="t5" @relatedToEventInstance="ei28"/>
<TLINK @relTypeA="vague" @relTypeB="before" @relTypeC="contains" @task="E2E"
eventInstanceID="ei26" @relatedToEventInstance="ei27"/>
<TLINK @relTypeA="before" @relTypeB="before" @relTypeC="before" @task="E2E"
eventInstanceID="ei27" @relatedToEventInstance="ei28"/>
```

English Translation:

Municipal Transportation Works Association published a report on April 10th. The report shows that only 4 public tube railways (e.g. Sapporo City Nanboku line) from 26 have a surplus.

Figure 1. An example of Japanese BCCWJ TimeBank annotation.

from the Japanese temporal information annotation. Then, we perform a comparison with the previous research on Asian languages.

3.1 Overview of Japanese Temporal Information Annotation

This section presents an overview of the Japanese temporal information annotation.

We introduce three types of tags for the annotation: i) the time expression tag <TIMEX3>, ii) the event expression and event instance tags <EVENT> and <MAKEINSTANCE>, iii) the temporal relation tag <TLINK>. We explain the specifications of each tag in the following subsections.

Note that our research objective is not the localization of ISO-TimeML for Japanese but is to investigate the ability of the human recognition process of temporal ordering. We omitted detailed localization of ISO-TimeML tags and attributes. In the original TimeML, <SIGNAL>, <SLINK>, and <ALINK> are defined. <SIGNAL> is used with some temporal prepositions and conjunctions in English, <SLINK> is used for subordination relations, and <ALINK> is used for non-constituent aspectual relations. Currently, we are not using these with the BCCWJ-TimeBank. We leave these directions for our future work for the following reason. According to the specification <SIGNAL> in TimeML, ISO-TimeML, and (Setzer, 2001), we have to annotate nearly all of the functional words (auxiliary verbs and postpositions) that should be tagged <SIGNAL>. If we do not introduce any subcategories for each functional word, annotating all of the functional words is the same as annotating none of the functional words. <SLINK> should be annotated after the annotation of the subordinate clause boundaries with the clause functions. Although <ALINK> should be annotated as an aspectual compound main verb structure in Japanese, verbs other than the main verb in a compound verb are defined as auxiliary verbs, according to the definition of the UniDic POS tagset. As presented in Section 2.3, the previous work on Chinese also focuses on <TLINK>.

3.2 <TIMEX3>: Temporal Expression Annotation

The target temporal expressions for <TIMEX3> are DATE, TIME, DURATION, and SET by @type. We do not permit any nests of <TIMEX3>. We clip the expressions according to characters because Japanese does not have word delimitation spaces.

The attributes of @tid, @type, @value, @freq, @quant, and @mod have been inherited from the original TimeML. There is an issue regarding which calendar to use in porting TimeML to Japanese. In Japan, we use not only the Western calendar but also a native Japanese calendar that is based on the year of the Emperor’s reign. We introduce a new attribute @valueFromSurface to address this issue. @valueFromSurface includes a @value-like string to indicate a machine-readable date/time value. @value includes the normalised version of value, whereas @valueFromSurface includes the non-normalised version of the value, which can be generated on rewrite rules. @valueFromSurface can encode Japanese calendars. For example, ‘平成26年’ (the 26th year of the Heisei era) is encoded in the @valueFromSurface as ‘H26’ and normalised as the @value of ‘2014’ in the ISO-8601-like format.

The difference between @value and @valueFromSurface shows the use of the normalisation procedure. Nevertheless, we cannot judge whether the <TIMEX3> is fully normalized (fully specified) or underspecified. We introduce another new attribute, @definite, to indicate whether the <TIMEX3> is fully specified ‘true’ or underspecified ‘false’.

3.3 <EVENT>, <MAKEINSTANCE>: Event Expression Annotation

Next, we need to annotate the event expressions and instances to link the temporal ordering to <TIMEX3>. The event expression candidates are automatically extracted from the BCCWJ of morphological information. We define *long word units* with verbs and adjectives, resulting in a total of 4,953 event expression candidates. First, the candidates are judged by two annotators as to whether the target expression is an event expression. If the expression boundaries are not valid, a longer expression covering the candidate is redefined by the annotators. Second, the annotators judge whether there are any instances of the target expression on the timeline. If an instance is recognised, the annotators define a <MAKEINSTANCE> in the corpus. The <MAKEINSTANCE> is a stand-off from the event expression, but is linked to the <EVENT> tag by the @eid attribute. Third, the annotators annotate the @class attribute on the <EVENT>. There are nine @class attributes: seven for event instances (OCCURRENCE, REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, I_STATE, STATE) and two for non-instances (NULL and NONE). The difference between NULL and NONE is that the former is applied by <EVENT> annotators and the latter by <TLINK> annotators. The instances are double-checked by both <EVENT> and <TLINK> annotators. These attributes are described in more detail below.

- OCCURRENCE: These are event expressions without event arguments describing something that happens or occurs in the world (the argument event). Most event expressions belong to this class.
- REPORTING: These are event expressions with an event argument describing the action of an animate actor declaring, narrating, or informing about the argument event.
- PERCEPTION: These are event expressions with an event argument describing the physical perception of the argument event.
- ASPECTUAL: These are event expressions with an event argument describing some aspectual feature of the argument event.
- I_ACTION: These are intentional action expressions with an event argument describing an action or situation to introduce the argument event, from which we can infer something, given its relation with the I_ACTION.
- I_STATE: These are intentional state expressions with an event argument referring to an alternative or possible world.
- STATE: These are state expressions in the timeline. We only annotate these when an instance is introduced and becomes an argument of other event expressions.
- NULL, NONE: These are non-instance expressions

The annotator discriminates whether the target is an event or a state (STATE). Then, he or she judges whether the target has an event argument (OCCURRENCE). Finally, he or she categorises any target with an event argument into one of the five categories of REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, and I_STATE.

The two annotators and two supervisors defined a detailed linguistic annotation specification employing some Japanese language tests that are based on linguistic research (Kudo, 1995, 2004; Nakamura, 2001). The two annotators were trained on the specification until the agreement rate reached 75%.

<MAKEINSTANCE> attributes such as @tense, @aspect, and @modality are not well-maintained in the current status. Japanese tense and aspect from the surface forms cannot be matched to ISO-TimeML. Japanese 'surface tense' is marked by *-u* and *-ta* and 'surface aspect' is marked by *-u* and *-teiru*.

In the case of tense, the surface tense only expresses the difference between past and non-past. Furthermore, the event marked by the surface tense may have discrepancies with the temporal ordering on the timeline. Nevertheless, the temporal ordering relation <TLINK> between DCT and the target event expresses 'deep tense' information. The deep tense information is a translingual feature. When we want to investigate the discrepancy between surface tense and deep tense, it can be done by comparing tense morphemes from the *short word unit* and <TLINK> with DCT.

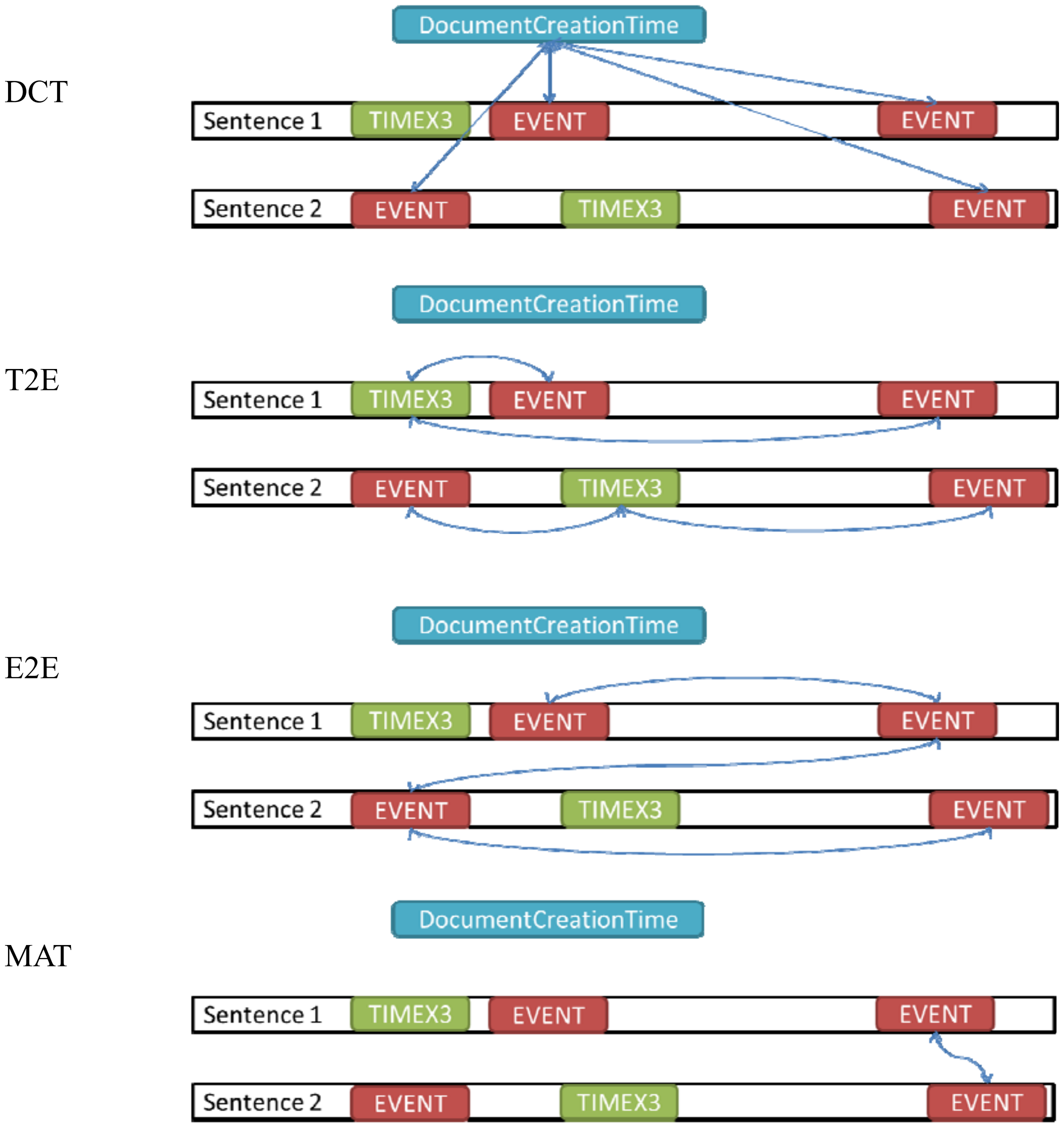
In the case of aspect, Japanese has some alternative surface aspectual expressions such as *-tearu* and *-teoru*. These aspectual expressions, excluding *-ru* and *-teiru*, are called 'semi-aspectual expressions' in Kudo (1995). These rich surface aspectual systems cannot be mapped on the ISO-TimeML original labels. Furthermore, the aspectual systems in Japanese vary among regions or dialects (Kudo, 2004). We believe this issue is beyond the scope of our current work, and we will treat this issue in our future work.

In the case of modality, Yamanashi University and Tohoku University developed annotation of modality information in BCCWJ. The researchers focused on whether the target event happened or did not happen in the real world in their annotation schema. Their research objective is different from ours. Nevertheless, they provide rich information on tense aspect and modality structure in their research. Layering their annotation on BCCWJ-TimeBank makes the corpus more informative.

3.4 <TLINK>: Temporal Ordering Annotation

<TLINK> defines the temporal ordering of temporal information expressions and event expressions. We used a variant of Allen's interval algebra as <TLINK> labels; there are 13 labels for temporal ordering and three for event-subevent relations.

We also have one label 'vague' for underspecified relations. Figure 3 shows the thirteen labels for temporal ordering and the three for event-subevent relations. The three underlined labels, namely ‘is_included,’ ‘identity,’ and ‘includes,’ are event-subevent versions of ‘during,’ ‘equal,’ and ‘contains,’ respectively. Strictly speaking, we can also define event-subevent versions of ‘finishes,’ ‘started-by,’ ‘starts,’ and ‘finished-by’. We did not define these, however, because they are rare and TimeML did not define them.



Note that Japanese is a strictly head-final language.
The matrix verb phrases tend to be put near the end of sentence.

Figure 2. <TLINK>: The four types of relations.

<TLINK> annotators are different from <EVENT> and <MAKEINSTANCE> annotators. Three annotators annotate the <TLINK> labels on some of the pairs among <TIMEX3> and <MAKEINSTANCE>. The number of <TLINK> candidates is equal to the combination number of <TIMEX3> and <MAKEINSTANCE>. It is difficult to check all possible pairs in the documents; therefore, we limited the target pairs to the following four types of relations:

- ‘DCT’: relations between a <TIMEX3> of document creation time (DCT) and an event instance;
- ‘T2E’: relations between a <TIMEX3> (non DCT) and an event instance within one sentence;
- ‘E2E’: relations between two consecutive event instances; and
- ‘MAT’: relations between two consecutive matrix verbs of event instances.

Figure 2 presents the four types of relations.

If the relation is between two different possible worlds, we use the label ‘vague’. When we regard the ‘vague’ relations as disjointed links, the connected subgraph indicates the different possible worlds. The value of <TIMEX3> is regarded not as a time point but as a time interval. The event instance of a punctual verb is regarded as a time point occurrence, whereas the other event instances are regarded as time interval occurrences. Figure 3 presents an overview of the <TLINK> labels.

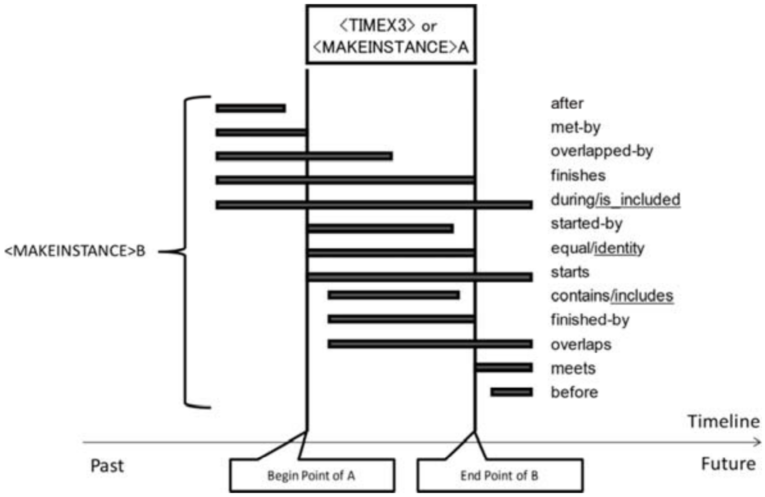


Figure 3. <TLINK> labels.

3.5 Other Tags in TimeML and ISO-TimeML

We will annotate <SIGNAL>, <SLINK>, and <ALINK> in the original TimeML tag in the future. In terms of the BCCWJ annotation background, several institutes are working together to take responsibility for layers of annotations. Some tags in TimeML originals presuppose other layers' annotation. To reduce annotation cost and to keep consistency among the professionals in other institutes, we postpone <SIGNAL>, <SLINK>, and <ALINK> until the presupposed annotation is finished. For example, <SIGNAL> tags are highly related to the multi-word functional words. <SLINK> tags are related to the subordinate clause structures in the treebank annotation. <ALINK> tags are related to the compound verb construction annotation.

3.6 Comparison with other Temporal Information Annotated Corpora in Asian Languages

This section presents a contrastive comparison with the previous research on Asian language resources.

First, the IREX NE Task data in Japanese (Sekine & Isahara, 2000) and the Chinese part of the ACE 2005 multilingual training corpus (Walker *et al*., 2006) focused only on time expression extraction and normalization.

Second, both of Cheng's works (Cheng *et al*., 2008a) and the TempEval-2 Chinese data sets focus on automatic <TLINK> annotation. <SLINK> and <ALINK> are not annotated on these resources. <TLINK> in Cheng's work is an annotated subset of the event and time expression pairs, and it expands the relation among the transition rules of temporal logic. On the other hand, <TLINK> in the TempEval-2 Chinese data sets targets limited pairs of the event and time expressions, which is almost the same as our approach, described in Section 3.4.

Third, KTimeML (Im *et al*., 2009) is a TimeML compatible annotation schema. Nevertheless, the researchers introduce several changes to the original TimeML markup philosophy, including a change from word-based in-line annotation to morpheme-based stand-off annotation. In the annotation definition, we inherit some <TIMEX3> standards from KTimeML, such as introduction of the week of the month.

Fourth, we discuss the word segmentation issue in Asian languages. CJK languages are written without word boundaries. The two Chinese resources are based on the Penn Chinese Treebank, and they use the word unit from the original Penn Chinese Treebank. KTimeML uses a morpheme unit from Korean. Nevertheless, they introduce the stand-off style annotation over the morpheme. BCCWJ-TimeBank uses the *short word unit* of the original BCCWJ.

There are still differences in several layers of the annotation among the Asian temporal information annotated language resources. Space limitations do not permit a full discussion of these differences.

4. Annotation Processes

This section presents the annotation processes. First, we present the MAMA-cycle and MATTER-cycle. Then, we introduce two additional paradigms for the annotation processes: MAMA-cycle with a pair-programming type of method for <TIMEX3> and annotation as a cognitive science experiment for <TLINK>. Note that we cannot introduce new annotation methods for <EVENT> and <MAKEINSTANCE>. We performed these two annotations by a simple MAMA-cycle.

4.1 MAMA-cycle and MATTER-cycle

O'Reilly's book, *Natural Language Annotation for Machine Learning* (Pustejovsky & Stubs, 2012), presents two types of annotation cycles. The MAMA-cycle, whose initials represent Model-Annotate-Model-Annotate, stresses the importance of creating a guideline or specification. The MATTER-cycle, whose initials represent Model-Annotate-Train-Test-Evaluate-Revise, stresses the importance of creating the language analysers. Train and Test are the phases of machine learning in creating language analysers. Figure 4 presents the two types of cycles.

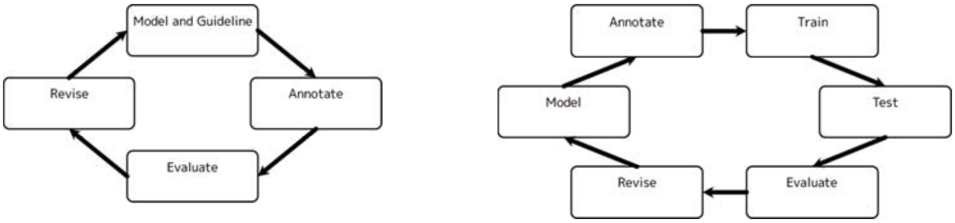


Figure 4. MAMA-cycle and MATTER-cycle.

The two cycles are aimed at producing effective guidelines, good language analysers, or both. Nevertheless, our research objective is different. The BCCWJ-TimeBank guidelines inherit most of the original TimeML/TimeBank schema. Our contribution to these guidelines is limited to the localization of the schema. We may produce language analysers in future work; however, we did not aim at the development of a temporal ordering relation analyser in the current stage. Our research objective is somewhat related to cognitive science because we would like to evaluate the human annotators' cognitive process of temporal ordering. Thus, in this work, we propose two additional annotation paradigms: MAMA-cycles with pair annotation and Annotation as a cognitive science experiment.

4.2 MAMA-cycle with Pair Annotation for <TIMEX3>

We used XML Editor oXygen 3 for <TIMEX3> annotation, and we defined DTD for BCCWJ-TimeBank. The DTD enables us to use the machine-aided (such as XML validation and a completion mechanism) environment of oXygen. An annotator performs in-line annotation on the original text corpus. We introduce a pair-programming type of method in which a display is shared by an annotator and supervisor. Although the method is stressful for both the annotator and supervisor, the data become more consistent and annotation errors are reduced. Figure 5 illustrates the proposed annotation process.

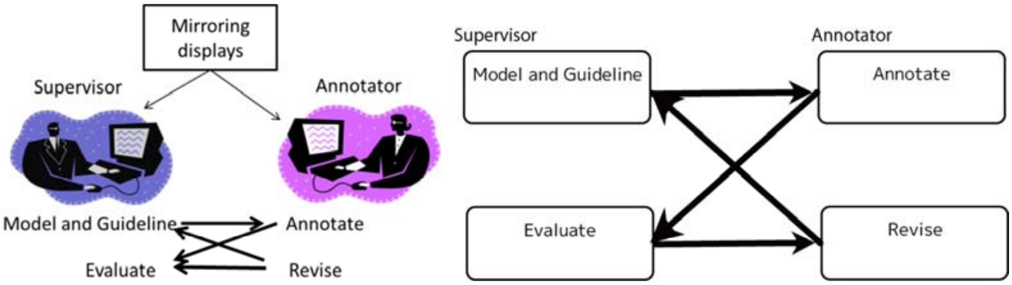


Figure 5. MAMA-cycle with a pair-programming method.

4.3 Annotation as a Cognitive Science Experiments for <TLINK>

Next, we performed a cognitive science experiment for <TLINK> annotations. In this paradigm, we evaluated the human cognitive process for perceiving the timeline. Here, we explain the process in detail. First, the supervisor gives annotation guidelines to three annotators. Second, the three annotators individually annotate <TLINK> information on the same dataset. Finally, the researchers evaluate the variance and differences among the three annotations. During this process, the annotators perform individual, not inter-annotator, revision. Figure 6 illustrates the proposed annotation process.

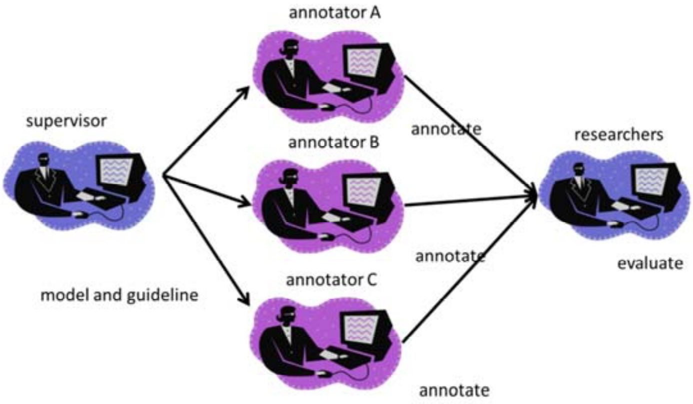


Figure 6. Annotation as a cognitive science experiment.

5. Corpus Statistics and Evaluations

This section presents basic statistics on BCCWJ-TimeBank, the Japanese corpus annotated for temporal information. We also consider the annotation environment of BCCWJ-TimeBank.

5.1 <TIMEX3>

Table 2 shows annotation target samples for <TIMEX3>. The column 'W/TIMEX' shows the number of samples or sentences that include at least one temporal information expression. Some samples in the registers OW (white paper), OC (Yahoo! Answers), and OY (Yahoo! Blogs) do not include any temporal information expressions.

Table 2. <TIMEX3>: Annotation target samples.

Register	# of Samples			# of Sentences			# of Words
(Priority set)	ALL	w/TIMEX	(Rate)	ALL	w/TIMEX		ALL
OW (A)	17	16	(94%)	1439	405	(28%)	58336
PB (A)	25	25	(100%)	2568	289	(11%)	57929
PN (A, B)	110	110	(100%)	5582	1562	(28%)	116834
OC (A)	518	250	(48%)	3479	488	(14%)	60086
PM (A)	23	23	(100%)	3066	413	(13%)	59372
OY (A)	257	198	(77%)	3986	765	(19%)	63459

Table 3 shows the basic statistics of <TIMEX3> annotations. The table shows the number of <TIMEX3> by @type and @definite and shows the relation of {@value and @valueFromSurface}. @type has four labels: DATE, TIME, DURATION, and SET. We exclude document creation time (DCT), which is given in corpus metadata, from the statistics. Then, we analyse the statistics on the basis of two perspectives. The first is whether @definite is 'true' or 'false,' in other words, whether the temporal information expression is fully specified or underspecified. The fully specified expression can be mapped on the timeline, while the underspecified expression cannot. The second perspective is whether @value and @valueFromSurface are identical ('=') or not ('≠'). The former have undergone some normalisation procedure from the annotators, while the latter have not.

A total of 5,297 temporal information expressions were annotated in the corpus. Of those, 1,639 (30%) are fully specified expressions without any normalisation procedures applied. Further, 2,023 (37%) of the total can be normalised by contextual information, and 1,875 (34%) cannot. The third group needs more external information to be able to be normalised.

Table 3. <TIMEX3>: @type×@definite×{@value, @valueFromSurface}.

@definite	True (fully specified)			False (underspecified)		
@value and @valueFromSurface	All	=	≠	All	=	≠
DATE	2214 (61%)	381 (10%)	1833 (50%)	1438 (39%)	1275 (35%)	163 (4%)
TIME	188 (37%)	1 (0%)	187 (37%)	315 (63%)	239 (48%)	76 (15%)
DURATION	1129 (92%)	1128 (92%)	1 (0%)	99 (8%)	99 (8%)	0
SET	131 (85%)	129 (84%)	2 (2%)	23 (15%)	22 (14%)	1 (1%)
ALL	3662 (66%)	1639 (30%)	2023 (37%)	1875 (34%)	1635 (30%)	240 (4%)

Count (Rate)

In the ‘DATE’ expressions, most of the fully specified expressions (@definite ‘true’; 61%) have had manual normalisation performed (@value ≠ @valueFromSurface; 50%). This fact shows that the normalisation procedure is important for temporal information processing. The normalisation includes conversion from the Japanese to the Western calendar, including, conversion from a 2-digit to a 4-digit calendar, and completion year (taken from the document creation time).

In the ‘TIME’ expressions, most fully specified expressions have had manual normalisation performed. The normalisation includes completion date (from the document creation time) and resolution of a.m./p.m. ambiguity.

In the ‘DURATION’ and ‘SET’ expressions, @definite ‘true’ means that the length of the temporal region on the timeline can be uniquely determined. When we map this on the timeline, we need <TLINK> information with ‘DATE’ or ‘TIME’ expressions or event expressions.

Note that we reduce the annotation target samples of <EVENT>, <MAKEINSTANCE>, and <TLINK> PN register (A), which total 54 samples. The reason is that only the PN (newspaper) sample has date-level document creation time information as metadata. Table 4 shows the statistics of <TIMEX3> in the PN (A) samples.

Table 4. <TIMEX3>: Statistics in PN (A).

DCT or class	Count
DCT(DATE)	54
DATE	727
TIME	107
DURATION	291
SET	19
ALL	1198

5.2 <EVENT> and <MAKEINSTANCE>

We annotate <EVENT> and <MAKEINSTANCE> tags only on the PN register (A). Table 5 shows the statistics of <EVENT> tags by @class. Event instances by <MAKEINSTANCE> are defined on the last seven @class in the tables. The number of <MAKEINSTANCE> is 3,839.

Table 5. <EVENT>: Statistics in PN (A).

<EVENT>@class	count
Non-instance	(1129)
NULL	1114
NONE	15
Event instance w/o event arg	(2352)
OCCURRENCE	2352
Event instance w/ event arg	(1291)
REPORTING	126
PERCEPTION	27
ASPECTUAL	63
I_ACTION	880
I_STATE	195
State instance	(181)
STATE	181

5.3 <TLINK>

The three annotators were independently trained for <TLINK> annotation. The annotation was performed on four types of relations: ‘DCT,’ ‘T2E,’ ‘E2E,’ and ‘MATRIX’.

Table 6 shows annotation agreement among the 13+3+1 labels by the three annotations and relation types. The three ∩-connected numbers are the label counts by each of the three annotators. The right number after ‘=’ is the agreed count.

The agreed counts for ‘after,’ ‘during,’ ‘contains,’ and ‘before’ are higher than the others. These relations do not exhibit boundary matching between the two time intervals. The relation ‘equal’ is the most frequent of those that do include interval boundary matching. Other relations are infrequent and show low agreement counts

across the three annotators. These findings reveal that a judgment of interval boundary matching is rare and is difficult for human annotators. The relation ‘vague’ was agreed on 314 times by the three annotators. This fact shows that the discrimination of possible worlds might be possible by annotation.

Table 6. <TLINK>: Annotation agreement by Annotators × Labels × Relation types.

Relation types	DCT	T2E	E2E	MATRIX	All
Count	3839	2188	2972	1245	10244
after	2352∩2326∩2133=1961	396∩441∩432=315	627∩631∩639=432	292∩284∩277=198	3667∩3682∩3481=2906
met-by	0∩0∩0=0	5∩10∩2=2	18∩12∩3=2	7∩3∩2=1	30∩23∩7=5
overlapped-by	11∩5∩4=2	59∩52∩42=20	3∩3∩2=0	0∩0∩1=0	73∩60∩49=22
finishes	2∩8∩1=0	10∩1∩11=0	5∩8∩5=1	1∩0∩0=0	18∩17∩17=1
during	449∩424∩650=217	105∩100∩113=62	206∩139∩225=67	112∩86∩134=43	872∩749∩1122=389
started-by	1∩0∩0=0	9∩2∩8=0	3∩14∩6=2	0∩3∩0=0	13∩19∩14=2
equal	1∩17∩0=0	37∩70∩51=19	263∩412∩307=154	62∩140∩90=29	363∩639∩448=202
starts	2∩0∩0=0	30∩9∩14=2	6∩16∩2=0	0∩1∩1=0	38∩26∩17=2
contains	164∩85∩144=63	830∩853∩868=671	299∩292∩344=117	148∩152∩188=64	1441∩1382∩1544=915
finished-by	0∩0∩0=0	3∩3∩0=0	6∩7∩6=0	1∩3∩0=0	10∩13∩6=0
overlaps	2∩2∩4=1	75∩84∩70=32	6∩27∩5=0	1∩4∩3=0	84∩117∩82=33
meets	1∩13∩0=0	25∩26∩2=2	88∩88∩32=22	9∩15∩0=0	123∩142∩34=24
before	739∩767∩746=572	389∩360∩383=288	1058∩994∩1098=713	418∩436=422=294	2604∩2557∩2649=1867
is_included	0∩0∩0=0	0∩0∩0=0	19∩2∩6=1	6∩0∩1=0	25∩2∩7=1
identity	0∩0∩0=0	0∩0∩1=0	11∩7∩24=2	16∩5∩15=2	27∩12∩40=4
includes	0∩0∩0=0	0∩0∩0=0	27∩10∩2=1	18∩2∩0=0	45∩12∩2=1
vague	115∩191∩157=38	212∩177∩191=100	327∩309∩265=128	154∩111∩111=48	808∩788∩724=314

Annotator A ∩ Annotator B ∩ Annotator C = Agreed count

Table 7 shows agreement rates by relation type across the three evaluation schemata. We define the schemata as follows. ‘Label 13+3+1’ is the most fine-grained evaluation schema; in it, all 13+3+1 relations are discriminated. ‘Label 13+1’ is a schema without event-subevent discrimination, in which ‘is_included,’ ‘identity,’ and ‘includes’ are regarded in the same way as ‘during,’ ‘equals,’ and ‘contains,’ respectively. ‘Label 5+1’ is a TempEval-like schema, in which 13+3+1 relations are generalised into 5+1 relations: ‘BEFORE,’ ‘BEFORE-OR-OVERLAP,’ ‘OVERLAP,’ ‘OVERLAP-OR-AFTER,’ ‘AFTER,’ and ‘VAGUE’.

The agreement rate across all relations is 65.3% (Cohen’s kappa 0.733) using the most fine-grained evaluation schema (Label 13+3+1). We perform <TLINK> annotations on fixed relation pairs for four types. TimeBank 1.2 jointly performs <TLINK> annotations without fixing the relation pairs. In this method, the <TLINK> relation agreement rate is 77% and the relation pair agreement 55%. We believe that the BCCWJ-TimeBank <TLINK> relation agreement rate is in no way inferior to that of TimeBank 1.2. Among the four relation types, the agreement rate of 'DCT' is the highest and that of 'T2E' is the second highest. The relation between a temporal information expression and an event instance is easier to determine than the relation between two event instances. This is because the interval of the temporal information expression is more easily defined on the timeline than the interval of the event instance is. Under the relaxed relation evaluation schema, the agreement rates of 'E2E' and 'MATRIX' increase. This means that, while interval boundary matching in these event instances is hard for the annotators to agree upon, interval anteroposterior relations can be agreed upon.

Table 7. <TLINK>: Annotation agreement by relation type across three evaluation schemata.

Relation types	DCT	T2E	E2E	MATRIX	ALL
Total Count	3839	2188	2972	1245	10244
Label 13+3+1	0.743(2854)	0.691(1513)	0.552(1642)	0.545(679)	0.653(6688)
Label 13+1	0.743(2854)	0.691(1513)	0.561(1667)	0.560(697)	0.657(6731)
Label 5+1	0.747(2873)	0.734(1605)	0.627(1862)	0.623(776)	0.695(7116)

Agreement rate(Agreed count)

Finally, Table 8 shows agreement by two entity types: DCT and TIMEX of <EVENT>@class. Relations with STATE tend to show low agreement rates, and relations between DCT/TIMEX and STATE are lower than relations between DCT/TIMEX and other event instances. This is because recognition of the time interval boundaries of state expressions is difficult for the annotators. In event instances with event arguments, relations with REPORTING and I_ACTION tend to show higher than average agreement rates.

Table 8. <TLINK>: Annotation agreement by {DCT, <TIMEX3>, <EVENT>@class} × <EVENT>@class.

	DCT	TIMEX	OCC	REP	PER	ASP	I_A	I_S	STA	ALL
OCCURRENCE	0.739	0.702	0.551	0.625	0.286	0.718	0.559	0.592	0.422	0.656
Abbr. OCC	(2352)	(1320)	(1602)	(104)	(7)	(39)	(494)	(130)	(102)	(6159)
REPORTING	0.881	0.697	0.663	0.222	1.000	0.667	0.519	0.368	0.500	0.694
Abbr. REP	(126)	(66)	(95)	(9)	(2)	(3)	(52)	(19)	(12)	(385)
PERCEPTION	0.815	0.700	0.444	NaN	0.000	NaN	0.500	1.000	0.000	0.646
Abbr. PER	(27)	(10)	(18)	(0)	(1)	(0)	(6)	(1)	(1)	(65)
ASPECTUAL	0.714	0.615	0.545	1.000	0.000	0.000	0.643	0.000	0.000	0.627
Abbr. ASP	(63)	(52)	(44)	(6)	(2)	(2)	(14)	(1)	(1)	(185)
I_ACTION	0.808	0.720	0.576	0.690	0.667	0.765	0.631	0.527	0.333	0.698
Abbr. I_A	(880)	(567)	(491)	(29)	(6)	(17)	(309)	(55)	(51)	(2407)
I_STATE	0.651	0.686	0.490	0.250	0.750	0.429	0.545	0.875	0.333	0.594
Abbr. I_S	(195)	(86)	(145)	(4)	(4)	(7)	(55)	(16)	(15)	(527)
STATE	0.492	0.398	0.356	0.600	1.000	0.444	0.431	0.333	0.238	0.424
Abbr. STA	(181)	(83)	(118)	(5)	(3)	(9)	(51)	(9)	(21)	(481)
ALL	0.743	0.691	0.548	0.618	0.560	0.649	0.573	0.562	0.374	0.653
	(3839)	(2188)	(2524)	(157)	(25)	(77)	(984)	(233)	(203)	(10244)

Agreement rate
(Agreed count)

6. Conclusion

This paper has presented temporal information annotation on BCCWJ. This is the first corpus-based study on Japanese temporal information annotation with the ISO-TimeML standard. We adapted the temporal information annotation specification of the original TimeML and ISO-TimeML to the Japanese language in several layers: <TIMEX3>, <EVENT>, <MAKEINSTANCE>, and <TLINK>. In addition, we introduced two annotation paradigms for linguistic research on Japanese temporal information: the MAMA-cycle annotation with a pair-programming method on <TIMEX3> and the annotation as a cognitive experiment on <TLINK>. Finally, we evaluated the cognitive process in human temporal information processing. The achieved temporal ordering agreement rates were 65.3%.

In future research, we will continue to investigate machine-learning-based temporal ordering estimation. In English temporal ordering, the tense and aspect information in <MAKEINSTANCE> are important features. In Japanese temporal ordering, however, the morphologically overt information is ‘る(-ru)’ vs. ‘た(-ta)’ for non-past and past tense, and ‘る(-ru)’ vs. ‘ている(-teiru)’ for limited aspect. We will report the results of this temporal ordering estimation in future publications.

We also intend to take advantage of BCCWJ’s status as the first balanced large-scale shared corpus of Japanese by analysing our annotation in comparison to the syntactic and semantic annotations conducted on BCCWJ by several Japanese research institutes, as mentioned in Section 2.2. Since Japanese is a modality-rich language, the modality annotations by these other institutes will be important for temporal ordering.

Furthermore, we will continue to evaluate the cognitive process in human temporal information processing. In this study, the annotators were professionally trained. In our next study, we will use crowdsourcing with 200 experimental subjects for the temporal ordering annotation. We will compare the annotation results between trained annotators and crowdsourcing subjects and will evaluate any differences between the specialists and non-specialists to further our understanding of human temporal information processing.

Reference

Boguraev, B., & Kubota Ando, R. (2005). TimeML-compliant text analysis for temporal reasoning. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, 997-1003.

Cheng, Y., Asahara, M., & Matsumoto, Y. (2008a). Constructing a temporal relation tagged corpus of Chinese on dependency structure analysis. *Journal of Association for Computational Linguistics and Chinese Language Processing (CLCLP)*, *13*(2), 171-196.

Cheng, Y., Asahara, M., & Matsumoto, Y. (2008b). The effect of event type classification for temporal relation identification in Chinese text. *The International Journal of Computer Processing of Oriental Language*, *21*(2), 151-167.

Den, Y., Nakamura, J., Ogiso, T., & Ogura, H. (2008). A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the 6th Language Resources and Evaluation Conference(LREC-2008)*, 1019-1024.

Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A brief history. In *Proceedings of Volume 1: The 16th International Conference on Computational Linguistics (COLING-96)*, 466-471.

Im, S., You, H., Jang, H., Nam, S., & Shin, H. (2009). KTimeML: Specification of temporal and event expressions in Korean text. In *Proceedings of the 7th Workshop on Asian Language Resources (ACL-IJCNLP 2009)*, 115-122.

Kudo, M. (1995). Asupekuto, Tensu-Taikei to Tekusuto --- Gendai Nihongo no Jikan-no Hyougen (in Japanese). ‘Aspect, tense and text, time expression in contemporary Japanese’, Hituzi Syobo.

Kudo, M. (2004). Nihongo-no Asupekuto, Tensu, Muudo-Taikei Hyouujun-go Kenkyu-wo Koete (in Japanese). ‘Aspect, tense, mood of Japanese, beyond the standard Japanese’, Hituzi Syobo.

Li, H., Strötgen, J., Zell, J., & Gertz, M. (2014). Chinese temporal tagging with HeidelTime. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2014)*, 133-177.

Maekawa, K. (2008). Balanced corpus of contemporary written Japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, 101-102. Hyderabad, India. Association for Computational Linguistics.

Mani, I. (2006). Machine learning of temporal relations. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL-2006*), 753-760. Sydney, Australia. Association for Computational Linguistics.

Nakamura, C. (2001). Nihongo-no Jikan Hyougen (in Japanese). ‘Time expressions in Japanese’, Kuroshio Shuppan.

Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., *et al.*, (2003a). TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, 337-353. Uppsala, Sweden. Association for Computational Linguistics.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., *et al.*, (2003b). The TimeBank corpus. In *Proceedings of Corpus Linguistics 2003*, 647-656. Lancaster, UK. UCREL technical paper (number 16).

Pustejovsky, J., Lee, K., Bunt, H., & Romary, L. (2010). ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC-2010)*, 394-397. Valletta, Malta. European Language Resources Association.

Pustejovsky, J., & Stubs, A. (2012). *Natural Language Annotation for Machine Learning*, O’Reilly.

Sekine, S., & Isahara, H. (2000). IREX: IR and IE evaluation project in Japanese. In *The Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, 1475-1480. Athens, Greece. European Language Resources Association.

Sekine, S., Sudo, K., & Nobata, C. (2002) Extended named entity hierarchy. In *The Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. 1818-1824. Las Palmas, Canary Islands, Spain. European Language Resources Association.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., & Pustejovsky. J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 1-9. Atlanta, Georgia, USA, Association for Computational Linguistics.

Verhagen, M., Gaizaukas, R., Schilder, F., Hepple, M., Katz, G., & Pustejovsky, J. (2007). Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on SemEval-2007*, 75-80. Prague, Czech Republic, Association for Computational Linguistics.

Verhagen, M., Saurí, R., Caselli, T., & Pustejovsky, J. (2010). SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, 57-62. Stroudsburg, PA, USA, Association for Computational Linguistics.

Warker, C., Strassel, S., Medero, J., & Maeda, K. (2006). ACE 2005 Multilingual Training Corpus. *Linguistic Data Consortium*, Philadelphia.

Xue, N., Xia, F., Chiou, F-D., & Parmer, M. (2005). The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, *11*(2), 207-238.

Xue, N., & Zhou, Y. (2010). Applying syntactic, semantic and discourse constraints in Chinese temporal annotation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*: Poster Volume, 1363-1372.

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 3, September 2014, pp. 25-38

Transliteration Extraction from Classical Chinese Buddhist Literature Using Conditional Random Fields with Language Models

Yu-Chun Wang[*] , Karol Chia-Tien Chang[+] ,

Richard Tzong-Han Tsai[#] , and Jieh Hsiang[*]

Abstract

Extracting plausible transliterations from historical literature is a key issue in historical linguistics and other research fields. In Chinese historical literature, the characters used to transliterate the same loanword may vary because of different translation eras or different Chinese language preferences among translators. To assist historical linguists and digital humanities researchers, this paper proposes a transliteration extraction method based on the conditional random field method with features based on the language models and the characteristics of the Chinese characters used in transliterations. To evaluate our method, we compiled an evaluation set from two Buddhist texts, the Samyuktagama and the Lotus Sutra. We also constructed a baseline approach with a suffix array based extraction method and phonetic similarity measurement. Our method significantly outperforms the baseline approach, and the method achieves recall of 0.9561 and precision of 0.9444. The results show our method is very effective for extracting transliterations in classical Chinese texts.

Keywords: Ttransliteration Extraction, Classical Chinese, Buddhist Literation, Langauge Model, Conditional Random Fields, CRF.

[*] Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
E-mail: d97023@csie.ntu.edu.tw; jhsiang@ntu.edu.tw

[+] Department of Computer Science and Engineering, Yuan Ze University, Taiwan
E-mail: s1003325@mail.yzu.edu.tw

[#] Department of Computer Science and Information Engineering, National Central University, Taiwan
E-mail: thtsai@csie.ncu.edu.tw
The author for correspondence is Richard Tzong-Han Tsai.

1. Introduction

Cognates and loanwords play important roles in the research of language origins and cultural interchange. Therefore, extracting plausible cognates or loanwords from historical literature is a key issue in historical linguistics. The adoption of loanwords from other languages is usually through transliteration. In Chinese historical literature, the characters used to transliterate the same loanword may vary because of different translation eras or different Chinese language/dialect preferences among translators. For example, in classical Chinese Buddhist scriptures, the translation process of Buddhist scriptures from Sanskrit to classical Chinese occurred mainly from the 1st century to 10th century. In these works, the same Sanskrit words may be transliterated into different Chinese loanword forms. For instance, the surname of the Buddha, Gautama, is transliterated into several different forms, such as “瞿曇” (qü-tan) or “喬答摩” (qiao-da-mo), and the name “Culapanthaka” has several different Chinese transliterations, such as “朱利槃特” (zhu-li-pan-te) and “周利槃陀伽” (zhou-li-pan-tuo-qie). In order to assist researchers in historical linguistics and other digital humanities research fields, an approach to extract transliterations in classical Chinese texts is necessary.

Many transliteration extraction methods require a bilingual parallel corpus or text documents containing two languages. For example, Sherif & Kondrak (2007) proposed a method for learning the string distance measurement function from a sentence-aligned English-Arabic parallel corpus to extract transliteration pairs. Kuo *et al.* (2007) proposed a transliteration pair extraction method using a phonetic similarity model. Their approach is based on the general rule that, when a new English term is transliterated into Chinese (in modern Chinese texts, *e.g.* newswire), the English source term usually appears alongside the transliteration. To exploit this pattern, they identify all of the English terms in a Chinese text and measure the phonetic similarity between those English terms and their surrounding Chinese terms, treating the pairs with the highest similarity as the true transliteration pairs. Despite its high accuracy, this approach cannot be applied to transliteration extraction in classical Chinese literature since the prerequisite (of the source terms alongside the transliteration) does not apply.

Some researchers have tried to extract transliterations from a single language corpus. Oh & Choi (2003) proposed a Korean transliteration identification method using a Hidden Markov Model (HMM) (Rabiner, 1989). They transformed the transliteration identification problem into a sequential tagging problem in which each Korean syllable block in a Korean sentence is tagged as either belonging to a transliteration or not. They compiled a human-tagged Korean corpus to train a hidden Markov model with predefined phonetic features to extract transliteration terms from sentences by sequential tagging. Goldberg & Elhadad (2008) proposed an unsupervised Hebrew transliteration extraction method. They adopted an English-Hebrew phoneme mapping table to convert the English terms in a named entity lexicon into all of the possible Hebrew transliteration forms. The Hebrew transliterations then were used to train a Hebrew transliteration identification model. Nevertheless, Korean and Hebrew use an alphabetical writing system, while Chinese is ideographic. These identification methods heavily depend on the phonetic characteristics of the writing system. Since Chinese characters do not necessarily reflect actual pronunciation, these methods are difficult to apply to the transliteration extraction problem in classical Chinese.

This paper proposes an approach to extract transliterations automatically in classical Chinese texts, especially Buddhist scriptures, with supervised learning models based on the probability of the characters used in transliterations and the language model features of Chinese characters.

2. Method

To extract the transliterations from the classical Chinese Buddhist scriptures, we adopted a supervised learning method, the conditional random fields (CRF) model. The features we used in the CRF model are described in the following subsections.

2.1 Probability of each Chinese Character in Transliterations

According to our observation, in the classical Chinese Buddhist texts, the Chinese characters used in transliteration show some characteristics. Translators were inclined to choose characters without obstructing the comprehension of the sentences. Although the number of Chinese characters is large, the number of possible syllables in Chinese is limited. Therefore, one Chinese character may share the same pronunciation with several other characters, and a translator may choose rarely used characters for transliteration.

Thus, the probability of a Chinese character being used in transliteration is an important feature to identify transliteration in the classical Buddhist texts. In order to measure the probability of every character used in transliterations, we collected the frequency of all the Chinese characters in the Chinese Buddhist Canon. Then, we applied the suffix array method (Manzini & Ferragina, 2004) to extract the terms with their counts from all the texts of the Chinese Buddhist Canon. The extracted terms then were filtered through a list of selected transliteration terms from the Buddhist Translation Lexicon and Ding Fubao’s Dictionary of Buddhist Studies. The extracted terms in the list were retained, and the frequency of each Chinese character was calculated. Thus, the probability of a given Chinese character c in transliteration can be defined as:

$$Prob(c) = \log \frac{freq_{trans}(c)}{freq_{all}(c)} \tag{1}$$

where $freq_{trans}(c)$ is c's frequency used in transliterations, and $freq_{all}(c)$ is c's frequency appearing in the entire Chinese Buddhist Canon. The logarithm in the formula is designed for CRF discrete feature values.

2.2 Character-based Language Model of the Transliteration

Transliterations may appear many times in one Buddhist sutra. The preceding character and the following character of the transliteration may be different. For example, for the phrase "於憍薩羅國" (yu-jiao-sa-luo-guo, "in Kosala state"), if we want to identify the actual transliteration, "憍薩羅" (jiao-sa-luo, Kosala), from the extra characters "於" (yu, in) and "國" (guo, state), we must first use an effective feature to identify the boundaries of the transliteration.

In order to do that, we propose a language-model-based feature. A language model assigns a probability to a sequence of m words $P(w_1,w_2,...,w_m)$ by means of a probability distribution. The probability of a sequence of m words can be transformed into a conditional probability:

$$\begin{aligned} P(w_1, w_2, \ldots, w_m) &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)\ldots P(w_m \mid w_1, w_2, \ldots w_{m-1}) \\ &= \prod_{i=1}^{m} P(w_i \mid w_1, w_2, \cdots, w_{i-1}) \end{aligned} \tag{2}$$

In practice, we can assume that the probability of a word only depends on its previous word (bi-gram assumption). Therefore, the probability of a sequence can be approximated as:

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, w_2, \cdots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-1}) \tag{3}$$

We collected person and location names from the Buddhist Authority Database and the known Buddhist transliteration terms from The Buddhist Translation Lexicon (翻譯名義集)[1] to create a dataset with 4,301 transliterations for our bi-gram language model. We used these transliterations to train the bi-gram language model. Such a language model may suffer from the sparse data problem. Nevertheless, since we adopted the language models as a feature for a supervised learning model, the sparse data problem is not serious in our approach.

After building the bi-gram language model, we applied it as a feature for the supervised model. Following the previous example, "於憍薩羅國" (yu-jiao-sa-luo-guo, "in Kosala state"), for each character in the sentence, we first computed the probability of the current

[1] http://www.cbeta.org/result/T54/T54n2131.htm

character and its previous character. For the first character “於”, since there is no previous word, the probability is P(於). For the second character “憍”, the probability of the two characters is P(於憍) = P(於)P(憍|於). We then computed the probability of the second and third characters: P(憍薩) = P(憍)P(薩|憍), and so on. If the probability changes sharply from that of the previous bi-gram, the previous bi-gram may be the boundary of the transliteration. Since the character “於” rarely appears in transliterations, P(於憍) is much lower than P(憍薩). We may conclude that the left boundary is between the first two characters “於憍”.

2.3 Pronunciation-based Language Model of the Transliteration

In addition to the character-based language model mentioned in the previous section, we also constructed a language model based on the pronunciations instead of characters. Since many Chinese characters may have the same pronunciation, different Chinese characters might be chosen to translate the same Sanskrit term. For example, the Sanskrit term, Arhat, has different Chinese transliteration forms, such as “阿羅訶” (a-luo-he) and “阿羅呵” (a-luo-he). The last Chinese characters (“訶” and “呵”) are different, but the Chinese pronunciations are the same. Therefore, a language model based on the pronunciation instead of Chinese character may overcome this kind of character variation problem. In order to construct a pronunciation based language model, the Chinese characters have to be converted into phoneme forms.

The pronunciation of Chinese characters, however, varies diachronically and synchronically. The same Chinese characters may have been pronounced differently in different regions and eras of ancient China. Therefore, we cannot base our language model on modern Chinese pronunciation. Furthermore, Chinese characters are ideographic and the pronunciations may not be reflected by the ideogram. Thus, it is difficult to find out how a character was pronounced in the past. In the seventh century (in the Sui dynasty), a new kind of pronunciation dictionary, *Qieyun*, was published by Lu Fayan, based on five earlier rhyming dictionaries that no longer exist. As a guide to the recitation of literary texts and an aid in the composition of verse, *Qieyun* quickly became popular and became the national standard of pronunciation during the ensuing Tang dynasty (618 - 907 C.E.). Unfortunately, the actual content of the *Qieyun* did not last into the modern era. In 1008, during the Northern Song Dynasty (960 - 1127 C.E.), a group of scholars commissioned by the emperor produced an expanded revision of *Qieyun* called *Guangyun*. Until the mid-20th century, the oldest complete rhyming book known was *Guangyun*, although existing copies are marred by numerous transcription errors. Thus, all studies of the rhyming book tradition were actually based on *Guangyun*. Since the period of the Buddhist literature translation from Sanskrit to classical Chinese is mainly from Tang dynasty to Song dynasty, which all belong to middle

Chinese era, we use *Guangyun* as an approximation to the pronunciation of Chinese characters in middle Chinese.

Since there were no phonological symbols or alphabetical writing systems in middle Chinese, rhyming books like Guangyun record contemporary character pronunciations with fanqie “反切” analyses. Fanqie represents a character’s pronunciation with another two characters, combining the former’s “initial” and the latter’s “rhyme” and tone. An English equivalent would be to combine the initial of ‘peek’ /p^{h}i:k/ and the rhyme of ‘cat’ /kæt/ to get ‘pat’ /p^{h}æt/. Take the character “東” [tuŋ] for example. The fanqie of the character is “德紅” ([tok] and [huŋ]), so that we can get its pronunciation if we know the actual pronunciation of these two fanqie characters. Although fanqie is an effective method to represent the pronunciation of a Chinese character, it is still hard to analyze because the usage of two characters for the initial and rhyme is arbitrary. Fortunately, a revision of *Guangyun* included additional annotations by some scholars. They analyzed all the characters used in fanqie and categorized the homophones into groups and chose an identical Chinese character standing for the group. After the analysis, there are 36 initials and 106 rhymes in *Guangyun*. In addition to the initial and rhyme, there are other features added into the revision of Guangyun, such as fanqie, initial, rhyme, openness (round or unround), level (different medial vowels), and tone.

To employ the data from *Guangyun* in our analyses, we must first convert the Chinese characters it uses to represent pronunciation into International Phonetic Alphabet (IPA) notation. There are many researchers who have tried to reconstruct the actual pronunciations of the characters in middle Chinese. We use the reconstruction of middle Chinese pronunciation proposed by Wang Li for this task. Take the character “洪” for example. Its initial “匣” is converted to IPA [ɣ] and its rhyme “東” is converted to [uŋ], giving us a final reconstructed IPA phonemic form of [ɣuŋ].

All of the Chinese characters used to construct the language model are converted into middle Chinese IPA representations by *Guangyun*. Nevertheless, one Chinese character might have several pronunciations. The homographs create difficulty in converting the Chinese characters into their actual pronunciations. Nevertheless, there are few tools and resources for classical Chinese and middle Chinese to deal with this problem. Therefore, we use a heuristic method to determine the most used pronunciation for each Chinese character. We found that the *Kangxi Dictionary* (康熙字典) often gives the most used pronunciation first for each Chinese character. Therefore, if one Chinese character has several different fanqie pronunciations, we check the description of the character in the Kangxi Dictionary and find the first matched fanqie as the final pronunciation of the character. Take the character “解” for example. In *Guangyun*, the character “解” has two fanqie pronunciations: “佳買” and “胡買”. The description of the character “解” in Kangxi Dictionary is “【唐韻】

【正韻】佳買切【集韻】【韻會】舉嶰切，㕕皆上聲。【說文】判也。从刀判牛角。【莊子・養生主】庖丁解牛。【左傳・宣四年】宰夫解黿。【前漢・陳湯傳】支解人民。【註】謂解截其四支也。…”. We can find the first fanqie is “佳買”. Therefore, we can determine the most used pronunciation of the character “解” is “佳買”, then convert it into IPA representation form [kai].

The feature value of the pronunciation-based language model is similar to the character-based language model described in Section 2.2. Following the previous example, “於憍薩羅國” (in Kosala state), we first convert the characters in the sentence into IPA representation, such as [ʔo kiu sat la kuok]. We then compute the probability of the current character and its previous character. For the first character “於” [ʔo], since there is no previous word, the probability is P(ʔo). For the second character “憍” [kiu], the probability of the two characters is P(ʔo kiu) = P(ʔo)P(kiu|ʔo). We then compute the probability of the second and third characters: P(kiu sat) = P(kiu)P(sat|kiu), and so on.

2.4 Functional Words

We take classical Chinese functional words into consideration. These characters have special grammatical functions in classical Chinese and are seldom used to transliterate foreign names. This is a binary feature that records the character as a functional word or not. The functional words are listed as follows: 之 (zhi), 乎 (hu), 且 (qie), 矣 (yi), 邪 (ye), 於 (yu), 哉 (zai), 相 (xiang), 遂 (sui), 嗟 (jie), 與 (yu), and 噫 (yi).

2.5 Appellation and Quantifier Words

After observing the transliterations appearing in classical Chinese literature, we note that there are some specific patterns in the characters following the transliteration terms. Most of the characters following the transliteration are appellation or quantifier words, such as 山 (san, mountain), 海 (hai, sea), 國 (guo, state), 洲 (zhou, continent). Examples are 耆闍崛山 (qi-du-jui-san, Vulture mountain), 拘薩羅國 (jü-sa-luo-guo, Kosala state), and 瞻部洲 (zhan-bu-zhou, Jambu continent). Therefore, we collect the Chinese characters that usually are used as appellation or quantifiers following transliterations and design this feature. This is a binary feature that records whether a character is used as an appellation or quantifier word or not.

2.6 CRF Model Training

We adopted the supervised learning models, conditional random field (CRF) (Lafferty *et al.*, 2011), to extract the transliterations in classical Buddhist texts. For the CRF model, we formulated the transliteration extraction problem as a sequential tagging problem.

2.6.1 Conditional Random Fields

Conditional random fields (CRFs) are undirected graphical models trained to maximize a conditional probability (Lafferty *et al.*, 2011). A linear-chain CRF with parameters $\Lambda = \lambda_1, \lambda_2, \ldots$ defines a conditional probability for a state sequence $\mathbf{y} = \mathbf{y}_1 \ldots \mathbf{y}_T$, given that an input sequence $\mathbf{x} = \mathbf{x}_1 \ldots \mathbf{x}_T$ is

$$P_\Lambda(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{x}, t) \right) \tag{4}$$

where Z_x is the normalization factor that makes the probability of all state sequences sum to one, $f_k(\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{x}, t)$ is often a binary-valued feature function, and λ_k is its weight. The feature functions can measure any aspect of a state transition, $\mathbf{y}_{t-1} \rightarrow \mathbf{y}_t$, and the entire observation sequence, **x**, centered at the current time step, *t*. For example, one feature function might have the value 1 when $\mathbf{y}_{t-1}$ is the state B, $\mathbf{y}_t$ is the state I, and $\mathbf{x}_t$ is the character “國” (guo). Large positive values for λ_k indicate a preference for such an event; large negative values make the event unlikely.

The most probable label sequence for **x**,

$$\mathbf{y}^* = \arg\max_{y} P_\Lambda(\mathbf{y} \mid \mathbf{x}) \tag{5}$$

can be efficiently determined using the Viterbi algorithm.

2.6.2 Sequential Tagging and Feature Template

The classical Buddhist texts were separated into sentences by the Chinese punctuation. Then, each character in the sentences was taken as a data row for CRF model. We adopted the tagging approach motivated by the Chinese segmentation (Tsai *et al.*, 2006) which treats Chinese segmentation as a tagging problem. The characters in a sentence are tagged in **B** class if it is the first character of a transliteration word or in **I** class if it is in a transliteration word but not the first character. The characters that do not belong to a transliteration word are tagged in **O** class. We adopted the CRF++ open source toolkit[2]. We trained our CRF models with the unigram and bigram features over the input Chinese character sequences. The features are shown as follows.

Unigram: $s_{-2}, s_{-1}, s_0, s_1, s_2$

Bigram: $s_{-1}s_0, s_0s_1$

[2] http://crfpp.googlecode.com

where the current substring is s_0 and s_i is other characters relative to the position of the current character.

3. Evaluation

3.1 Data Set

We chose Samyuktagama (雜阿含經), a Buddhist scripture from the Chinese Buddhist Canon maintained by Chinese Buddhist Electronic Text Association (CBETA), as our data set for evaluation. The Samyuktagama is one of the most important scriptures in Early Buddhism and contains a lot of transliterations because it records in detail the teachings and the lives of the Buddha and many of his disciples.

The Samyuktagama is an early Buddhist scripture collected shortly after the Buddha's death. The term agama in Buddhism refers to a collection of discourses, and the name Samyuktagama means "connected discourses". It is among the most important sutras in Early Buddhism. The authorship of the Samyuktagama is traditionally attributed to Mahakssyapa, Buddha's disciple, and five hundred Arhats three months after the Buddha's death. An Indian monk, Gunabhadra, translated this sutra into classical Chinese in the Liu Song dynasty around 443 C.E. The classical Chinese Samyuktagama has 50 volumes containing about 660,000 characters. As the amount of text in the Samyuktagama is immense, we took the first 20 volumes as the training set, and the last 10 volumes as the test set.

We also wanted to see whether the supervised learning model trained by one Buddhist scripture can be applied to another Buddhist scripture translated in a different era. Therefore, we chose another scripture, the Lotus Sutra (妙法蓮華經), to create another test set. The Lotus sutra is a famous Mahayana Buddhist scripture probably written down between 100 BC and 100 C.E. The earliest known Sanskrit title for the sutra is the Saddharma Pundarika Sutra, which translates to "the Good Dharma Lotus Flower Sutra". In English, the shortened form Lotus Sutra is common. The Lotus Sutra has been regarded highly in a number of Asian countries where Mahayana Buddhism traditionally has been practiced, such as China, Japan, and Korea. The Lotus Sutra has several classical Chinese translation versions. The most widely used version was translated by Kumarajiva ("鳩摩羅什" in Chinese) in 406 C.E. It has eight volumes and 28 chapters containing more than 25,000 characters. We selected the first 5 chapters as a different test set to evaluate our method.

3.2 Baseline Method

There are a few research projects focused on transliteration extraction from classical Chinese literature. Nevertheless, in order to compare and show the effectiveness of our method, we constructed a baseline system with widely used information extraction methods. Since many previous research projects on transliteration extraction are based on phonetic similarity or phoneme mapping approaches, we also used these methods to construct the baseline system. First, the baseline system used the suffix array method to extract all the possible terms for the classical Chinese Buddhist scriptures. Then, the extracted terms were converted into Pinyin sequences by a modern Chinese pronunciation dictionary. We also adopted the collected transliteration list used in Section 2.1 and converted the transliterations into Pinyin sequences. Next, for each extracted term, the baseline system measured the Levenshtein distance between the Pinyin sequences of the extracted terms and all the transliterations as the phonetic similarity. If the extracted term had a Levenshtein distance less than the threshold (distance ≤ 3 in our baseline) from one of the transliterations we collected, the extracted term would be regarded as a transliteration; otherwise, the term would be dropped.

3.3 Evaluation Metrics

We used two evaluation metrics, recall and precision, to estimate the performance of our system. Recall and precision are widely used measurements in many research fields, such as information retrieval and information extraction (Manning *et al.*, 2008). In digital humanities, a key issue is the coverage of the extraction method. To maximize usefulness to researchers, a method should be able to extract as many potential transliterations from literature as possible. Therefore, in our evaluation, we used recall, defined as follows:

$$Recall = \frac{|\text{Correctly extracted transliterations}|}{|\text{Transliterations in the data set}|} \tag{6}$$

In addition, the correctness of the extracted transliterations is also important. To avoid wasting time on useless information, a method should be able to extract correct transliterations from literature as much as possible. Thus, we also used precision, defined as follows:

$$Precision = \frac{|\text{Correctly extracted transliterations}|}{|\text{All extracted transliterations}|} \tag{7}$$

With precision and recall, the F-score measurement also was adopted as a weighted average of the precision and recall. The F1-score is defined as follows:

$$F_1\text{-score} = \frac{2 \times precision \times recall}{precision + recall} \tag{8}$$

3.4 Evaluation Results

Table 1 shows the results of our method with two different language models and the baseline system on different test sets. The gold standards of these two test sets were compiled by human experts who examined all of the sentences in the test sets and recognized each transliteration for evaluation. The results show that our method with the character-based language model could extract 95.61% of the transliterations in the Sumyuktagama and 94.74% in the Lotus Sutra. On the precision measurement, our method also achieved pretty good results, which show that most of the terms our method extract are actual transliterations. The pronunciation-based language model does not perform well as the character-based one in both recall and precision metrics.

Table 1. Evaluation results of transliteration extraction.

	Data Set	Precision	Recall	F_1-score
Our Approach (character-based LM)	Samyuktagama	0.8810	0.9561	0.9170
	Lotus Sutra	0.9444	0.9474	0.9459
Our Approach (pronunciation-based LM)	Samyuktagama	0.8477	0.7530	0.7975
	Lotus Sutra	0.2081	0.6447	0.3146
Our Approach (character & pronunciation based LM)	Samyuktagama	0.8224	0.7349	0.7762
	Lotus Sutra	0.4581	0.7763	0.5762
Baseline	Samyuktagama	0.0399	0.7771	0.0759
	Lotus Sutra	0.0146	0.5789	0.2848

Our method outperforms the baseline system. The baseline system cannot extract most transliterations due to the limit of the suffix array method since the suffix array method only extracts the terms that appear twice or more often in the context. Furthermore, phonetic similarity is not effective to filter the transliterations, which causes the low precision performance of the baseline method. These results demonstrate that our method can save humanities researchers a lot of labor-intensive work in examining the transliteration.

4. Discussion

4.1 Effectiveness of Transliteration Extraction

Our method can extract many transliterations from the Samyuktagama, such as “迦毘羅衛” (jia-pi-luo-wei, *Kapilavastu*, the name of an ancient kingdom where the Buddha was born and raised), “尼拘律” (ni-jü-lü, *Nyagro*, the forest name in the Kapilavastu kingdom), and “摩伽陀” (muo-qie-tuo, *Magadha*, the name of an ancient Indian kingdom). These transliterations do not appear in the training set, but our method can still identify them. In addition, our method also discovered many transliterations in the Lotus Sutra that do not appear in the Samyuktagama, such as “娑伽羅” (suo-qie-luo, *Sagara*, the name of the king of the sea world in ancient Indian mythology), “鳩槃茶/鳩槃荼”

(jiu-pan-cha/jiu-pan-tu, *Kumbhanda*, one of a group of dwarfish, misshapen spirits among the lesser deities of Buddhist mythology), and “阿鞞跋致” (a-pi-ba-zhi, *Avaivart*, “not turn back” in Sanskrit). Since the characteristics of the Lotus Sutra are different from the Samyuktagama in many aspects, it shows that the supervised learning model trained by one Buddhist scripture may apply to other Buddhist scriptures translated in different eras and translators.

We have also discovered that transliterations may vary even in the same scripture. In the Samyuktagama, the Sanskrit term“Chandala” (someone who deals with the disposal of corpses and is a Hindu lower caste, formerly considered untouchable) has two different transliterations: “旃陀羅” (zhan-tuo-luo) and “栴陀羅” (zhan-tuo-luo).

The Sanskrit term “*Magadha*” (the name of an ancient Indian kingdom) has three different transliterations: “摩竭陀” (muo-jie-tuo), “摩竭提” (muo-jie-ti), and “摩伽陀” (muo-qie-tuo). The variations of the transliterations of the same word give clues of who the translators were and the progress of the translations. These variations may help the study of historical Chinese phonology and philology.

4.2 Comparison between Character-based and Pronunciation-based Language Models

From the evaluation results, we find that the pronunciation-based language model approach does not perform as well as the character-based one. Especially for the Lotus Sutra data set, the precision of the pronunciation-based approach drops sharply. Many non-transliteration candidates are extracted by the approach with the pronunciation-based language model, such as “逮得”, “何因”, “悅可”, “後必憂”, and “但離”. Since the pronunciation-based language model only considers the pronunciations instead of the actual characters and semantics, some terms that are not transliterations but have similar pronunciation patterns to those used in transliterations are extracted as false positives. The results also show that the supervised learning model with the pronunciation-based language model trained by the Samyuktagama does not predict well on other Buddhist literature, such as the Lotus Sutra. Since these two Buddhist works have many differences in content, the model that is only based on pronunciation cannot deal with the differences to get better results.

4.3 Error Cases

Although our method can extract and identify most transliteration pairs, some transliteration pairs cannot be identified. The error cases can be divided into several categories. The first one is that a few terms cannot be extracted, such as “闍維” (she-wei, *Jhapita*, cremation, a monk’s funeral pyre). This transliteration is seldom used and only appears three times in the final part of the Samyuktagama. The widely used transliteration of the term “*Jhapita*” is “荼

毘” (tu-pi). This may cause difficulty for the supervised learning model to identify these terms.

The other case is incorrect boundaries of the transliterations. Sometimes, our method may extract shorter terms, such as “韋提” (wei-ti, correct transliteration is “韋提希”, wei-ti-xi, *Vaidehi*, a female person name), “波羅” (po-luo, correct transliteration is “波羅柰”, po-luo-nai, *Varanasi*, a location name in Northern India), “瞿利摩羅” (qü-li-muo-luo, correct transliteration is “央瞿利摩羅”, yang-qü-li-muo-luo, *Angulimala*, one of the Buddha’s disciples). This problem is due to the probability generated by the language model. For example, the probability of the first two characters of the transliteration “央瞿利摩羅”, *P*(央瞿), is very low. This causes the CRF model to predict that the first character “央” (yang) does not belong to the transliteration. If more transliterations can be collected to build a better language model, this problem can be overcome.

In some cases, our method extracts longer terms, such as “阿那律陀夜” (a-na-lü-tuo-ye) while the correct transliteration is “阿那律陀”, (a-na-lü-tuo, *Aniruddha*, one of the Buddha’s closest disciples); and “兒富那婆藪” (er-fu-na-po-sou), while the correct transliteration is “富那婆藪” (fu-na-po-sou, *Punabbasu*, a kind of ghost in Buddhist mythology). In these cases, the preceding or following characters are often used in transliterations. There are cases where a transliteration is immediately followed by another transliteration. For example, our method extracts the term “闡陀舍利” (chan-tuo-she-li), which actually comprises two transliteration terms “闡陀” (chan-tuo, *Chanda*, one of the Buddhist’s disciples) and “舍利” (she-li, *Sarira*, Buddhist relics). It is difficult to separate them without any additional semantic clues. Although our method sometimes might extract incomplete transliterations with incorrect boundary, checking the boundary of a transliteration is not difficult for a human expert. Therefore, the extracted incorrect transliterations also have the benefit of helping humanities researchers quickly find and check plausible transliterations.

5. Conclusion

The transliteration extraction of foreign loanwords is an important task in research fields, such as historical linguistics and digital humanities. We propose an approach that can extract transliteration automatically from classical Chinese Buddhist scriptures. Our approach comprises the conditional random fields method with designed features that are suitable to identify transliteration characters based on language models and textual characteristics. The first feature is the probability of each Chinese character used in transliterations. The second feature is probability of the sequential bigram characters or phonemic representations measured by the language model method. In addition, functional words, appellation, and quantifier words are regarded as binary features. The transliteration extraction problem is formulated as a sequential tagging problem, and the CRF method is used to train a model to extract the transliterations from the input classical Chinese sentences. To evaluate our method, we constructed an evaluation set from the two Buddhist texts, the Samyuktagama and the Lotus Sutra, which were translated into Chinese in different eras. We also constructed a baseline system with a suffix array based extraction method and phonetic similarity measurement for comparison. The recall of our method achieved 0.9561 and the precision was 0.9444. The results show our method outperforms the baseline system and is effective for extracting transliterations from classical Chinese texts. Our method can find the transliterations among the immense classical literature to help many research fields, such as historical linguistics and philology.

Reference

Goldberg, Y., & Elhadad, M. (2008). Identification of transliterated foreign words in hebrew script. *Computational Linguistics and Intelligent Text Processing*.

Kuo, J.-S., Li, H., & Yang, Y.-K. (2007). A phonetic similarity model for automatic extraction of transliteration pairs. *ACM Trans. Asian Language Information Processing*, *6*(2).

Lafferty, J., McCallum, A., & Pereira, F. (2011). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 29th Internation Conference on Machine Learning (ICML)*, 282-289.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*, Cambridge University Press Cambridge.

Manzini, G., & Ferragina, P. (2004). Engineering a lightweight suffix array construction algorithm. *Algorithmica*, *40*(1), 33-50.

Oh, J., & Choi, K. (2003). A statistical model for automatic extraction of Korean transliterated foreign words. *International Journal of Computer Processing of Oriental Languages*, *16*(1), 41-62.

Rabiner, L. (1989). Tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE)*, 77.

Sherif, T., & Kondrak, G. (2007). Bootstrapping a stochastic transducer for Arabic-English transliteration extraction. In *Proceedings of Annual Meeting Asociation for Computational Linguistics*.

Tsai, R. T.-H., Hung, H.-C., Sung, C.-L., Dai, H.-J., & Hsu, W.-L. (2006). On closed task of chinese word segmentation:An improved CRF model coupled with character clustering and automatically generated template matching. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 134-137.

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 3, September 2014, pp. 39-54

Modeling Human Inference Process for Textual Entailment Recognition

Hen-Hsen Huang*, Kai-Chun Chang* and Hsin-Hsi Chen*

Abstract

To prepare an evaluation dataset for textual entailment (TE) recognition, human annotators label rich linguistic phenomena on text and hypothesis expressions. These phenomena illustrate implicit human inference process to determine the relations of given text-hypothesis pairs. This paper aims at understanding what human think in TE recognition process and modeling their thinking process to deal with this problem. At first, we analyze a labelled RTE-5 test set which has been annotated with 39 linguistic phenomena of 5 aspects by Mark Sammons *et al.*, and find that the negative entailment phenomena are very effective features for TE recognition. Then, a rule-based method and a machine learning method are proposed to extract this kind of phenomena from text-hypothesis pairs automatically. Though the systems with the machine-extracted knowledge cannot be comparable to the systems with human-labelled knowledge, they provide a new direction to think TE problems. We further annotate the negative entailment phenomena on Chinese text-hypothesis pairs in NTCIR-9 RITE-1 task, and conclude the same findings as that on the English RTE-5 datasets.

Keywords: Textual Entailment Recognition, Chinese Processing, Semantic.

1. Introduction

Textual Entailment (TE) is a directional relationship between pairs of text expressions, text (*T*) and hypothesis (*H*). Given a text pair *T* and *H*, if human would consider that the meaning of *H* is right by using the information of *T*, then we can infer *H* from *T* and say that *T* entails *H* (Dagan, Glickman, & Magnini, 2006). (S1) shows an example where *T* entails *H*.

* Department of Computer Science and Information Engineering, National Taiwan University
E-mail: {hhhuang, kcchang}@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

(S1) **T**: Norway’s most famous painting, ‘The Scream’ by Edvard Munch, was recovered Saturday, almost three months after it was stolen from an Oslo museum.
H: Edvard Munch painted ‘The Scream’.

Because such an inference is important in many applications (Androutsopoulos & Malakasiotis, 2010), the researches on textual entailment have attracted much attention in recent years. Recognizing Textual Entailment (RTE) (Bentivogli *et al.*, 2011), a series of evaluations on the developments of English TE recognition technologies, have been held seven times up to 2011. In the meanwhile, TE recognition technologies in other languages are also underway. The 9th NTCIR Workshop Meeting first introduced a TE task in Chinese and in Japanese called Recognizing Inference in Text (RITE-1) into the IR series evaluation (Shima *et al.*, 2011).

The overall accuracy is used as the only evaluation metric in most TE recognition tasks (Androutsopoulos & Malakasiotis, 2010). However, it is hard to examine the characteristics of a system when only considering its performance by accuracy. Sammons *et al.*, (2010) proposed an evaluation metric to examine the characteristics of a TE recognition system. They annotated text-hypothesis pairs selected from the RTE-5 test set with a series of linguistic phenomena required in the human inference process. When annotators assume that some linguistic phenomena appear in their inference process to determine whether T entails H, they would label the T-H pair with these phenomena. The RTE systems are evaluated by the new indicators, such as how many T-H pairs annotated with a particular phenomenon can be correctly recognized. The indicators can tell developers which systems are better to deal with T-H pairs with the appearance of which phenomenon. On the other hand, that would give developers a direction to enhance RTE systems.

For example, (S2) is an instance that matches the linguistic phenomena Exclusive Relation, and this phenomenon suggests *T* does not entail *H*. More than one argument of *H*, i.e., Venus Williams, Marion Bartoli, 2007, and Wimbledon Championships, appear in *T*, but the relation defeated in *H* contracts the relation triumphed in *T*.

(S2) **T**: Venus Williams triumphed over Marion Bartoli of France 6-4, 6-1 yesterday to win the Women's Singles event at the 2007 Wimbledon Championships. For the first time, an American and Frenchwoman were matched up to compete for the British women's singles title. A Wimbledon champion in 2000, 2001 and 2005, Williams was not the favorite to win the title again this year. Currently ranked 23rd in the world, she entered the tournament in the shadow of her sister, Serena Williams.

H: Venus Williams was defeated by Marion Bartoli at the 2007 Wimbledon Championships.

Such linguistic phenomena are thought as crucial in the human inference process by annotators. In the RITE-2 in the 10th NTCIR Workshop Meeting, some linguistic phenomena for TE in Japanese are reported in the unit task subtask (Watanabe *et al.*, 2013). In a similar manner, types of some linguistic phenomena in Chinese are consulted in the RITE-VAL task in the 11th NTCIR Workshop Meeting[1]. In this paper, we use this valuable resource from a different aspect. Instead of using the labelled linguistic phenomena in the evaluation of TE recognition, we aim at knowing the ultimate performance of TE recognition systems which embody human knowledge in the inference process. The experiments show five negative entailment phenomena may be strong features for TE recognition, and this finding confirms the previous study of Vanderwende *et al* (2006). Moreover, we propose a method to acquire the linguistic phenomena automatically and use them in TE recognition. Our method is evaluated on both the English RTE-5 dataset and the Chinese NTCIR-9 RITE-1 dataset. Experimental results show that our method achieves decent performances near the average performances of RTE-5 and NTCIR-9 RITE-1. Compared to the other methods incorporating a lot of features, only a tiny number of binary features are required by our methods.

This paper is organized as follows. In Section 2 we introduce the linguistic phenomena used by annotators in the inference process, do a series of analyses on the human annotated dataset released by Mark Sammons et al., and point out five significant negative entailment phenomena. Section 3 specifies the five negative entailment phenomena in detail, proposes a rule-based method and a machine learning method to extract them from T-H pairs automatically, and discuss their effects on TE recognition. In Section 4, we extend the methodology to the BC (binary class subtask) dataset distributed by NTCIR-9 RITE-1 task (Shima *et al*., 2011), annotate the dataset similar to the schema of Sammons *et al*. (2010), discuss if the negative entailment phenomena also appear in Chinese T-H pairs, and show their effects on TE in Chinese. Section 5 concludes the remarks.

2. Analyses of Human Inference Process in Textual Entailment

We regard the human annotated phenomena as features in recognizing the binary entailment relation between the given T-H pairs, i.e., ENTAILMENT and NO ENTAILMENT. Total 210 T-H pairs were chosen from the RTE-5 test set by Sammons *et al*. (2010), and total 39 linguistic phenomena divided into the following 5 aspects as follows, including knowledge domains, hypothesis structures, inference phenomena, negative entailment phenomena, and

[1] https://sites.google.com/site/ntcir11riteval/home-ct/task-guideline

knowledge resources, are annotated on the selected dataset. Table 1 summarizes the phenomena in the five aspects.

(a) **Knowledge Domains (Hypothesis Types)**: Each phenomenon in this aspect denotes whether the information in H belongs to the corresponding knowledge domain.

(b) **Hypothesis Structures**: Each phenomenon in this aspect denotes whether the H contains elements of the corresponding type.

(c) **Inference Phenomena**: Each phenomenon in this aspect indicates the corresponding linguistic phenomenon which is used to infer H from T.

(d) **Negative Entailment Phenomena**: Each phenomenon in this aspect is a pattern which may appear in negative entailment instances.

(e) **Knowledge Resources**: Each phenomenon in this aspect is a kind of knowledge or common senses which are required in the inference process in textual entailment.

Table 1. Five aspects of linguistic phenomena relating to textual entailment.

Aspect	Phenomena Types
Knowledge Domains	“be in”, “cause”, “come from”, “create”, “die/injure/kill”, “group”, “kinship”, “name”, “win/compete”, “work”
Hypothesis Structures	“has Named Entity”, “has Numerical Quantity”, “has implicit relation”, “has locative argument”, “has nominalization relation”, “has temporal argument”
Inference Phenomena	“coerced relation”, “co-reference”, “genitive relation”, “implicit relation”, “lexical relation”, “nominalization”, “passive-active”, “wrong-label”
Negative Entailment Phenomena	“Named Entity mismatch”, “Numeric Quantity mismatch”, “disconnected argument”, “disconnect relation”, “exclusive argument”, “exclusive relation”, “missing modifier”, “missing argument”, “missing relation”
Knowledge Resources	“event chain”, “factoid”, “parent-sibling”, “simple rewrite rule”, “spatial reasoning”, “numeric reasoning”

2.1 Five Aspects as Features

We train SVM classifiers to evaluate the performances of the five aspects of phenomena as features for TE recognition. The implementation LIBSVM with the RBF kernel (Chang & Lin, 2011) is adopted to develop classifiers with the parameters tuned by grid search. The experiments are done with 10-fold cross validation.

For the dataset of Sammons *et al.* (2010), two annotators are involved in labeling the above 39 linguistic phenomena on the T-H pairs. They may agree or disagree in the annotation. In the experiments, we consider the effects of their agreement. Table 2 shows the results. Five aspects are first regarded as individual features, and then merged together. The two schemes, *Annotator* 1 and *Annotator* 2, mean the phenomena labelled by annotator 1 and annotator 2 are used as features, respectively. The scheme "1 *AND* 2", a strict criterion, denotes a phenomenon exists in a T-H pair only if both annotators agree with its appearance. In contrast, the scheme "1 *OR* 2", a looser criterion, denotes a phenomenon exists in a T-H pair if at least one annotator marks its appearance.

We can see that the aspect of *negative entailment phenomena* is the most significant features of the five aspects. With only 9 phenomena in this aspect, the SVM classifier achieves accuracy above 90% no matter which labeling schemes are adopted. Comparatively, the best accuracy in RTE-5 task is 73.5% (Iftene & Moruz, 2009). In negative entailment phenomena aspect, the "1 *OR* 2" scheme achieves the best accuracy whereas the performances of Annotator 1 and "1 OR 2" are the same in the setting with all the five aspects as features. In the following experiments, we adopt this labeling scheme.

Table 2. The accuracy of recognizing binary TE relation with the five aspects as features.

Aspect	Annotator 1	Annotator 2	1 AND 2	1 OR 2
Knowledge Domains	50.95%	52.38%	52.38%	50.95%
Hypothesis Structures	50.95%	51.90%	50.95%	51.90%
Inference Phenomena	74.29%	72.38%	72.86%	74.76%
Negative Entailment Phenomena	97.14%	95.71%	92.38%	97.62%
Knowledge Resources	69.05%	69.52%	67.62%	69.52%
ALL	97.14%	92.20%	90.48%	97.14%

2.2 Negative Entailment Phenomena

There is a large gap between negative entailment phenomena aspect and the second effective aspect (i.e., inference phenomena). Moreover, using the negative entailment phenomena aspect as features only is even better than using all the 39 linguistic phenomena as features. We further analyze which negative entailment phenomena are more significant.

There are nine linguistic phenomena in the aspect of negative entailment phenomena. We take each phenomenon as a single feature to do the task of two-way textual entailment recognition. Table 3 shows the experimental results. The first column is the phenomenon ID, the second column is the phenomenon, and the third column is the accuracy of using the phenomenon in the binary classification. Comparing with the best accuracy 97.62% shown in Table 2, the highest accuracy in Table 3 is 69.52%, when missing argument is adopted. Each phenomenon may be suitable for some T-H pairs, and consequently all negative entailment phenomena together achieve the best performance.

Table 3. Accuracy of recognizing TE relation with individual negative entailment phenomena.

Phenomenon ID	Negative entailment Phenomenon	Accuracy
0	Named Entity mismatch	60.95%
1	Numeric Quantity mismatch	54.76%
2	Disconnected argument	55.24%
3	Disconnected relation	57.62%
4	Exclusive argument	61.90%
5	Exclusive relation	56.67%
6	Missing modifier	56.19%
7	Missing argument	69.52%
8	Missing relation	68.57%

We consider all possible combinations of these 9 negative entailment phenomena, i.e., C_1^9+…+ C_9^9 =511 feature settings, and use each feature setting to do the task of two-way entailment relation recognition by SVM classifiers. The notation C_n^m denotes a set of $m!/((m\text{-}n)!\times n!)$ feature settings, each with n features. For the sake of paper space, we only list the best 4 results in each combination set C_n^m shown in Table 4. Each feature setting is denoted by a set of phenomenon IDs enclosed parentheses. The notations between combination sets C_1^9~C_4^9 and C_5^9 ~C_8^9 are a slight difference because of the table space. For clarification, we list the phenomena not involved in the combination sets C_5^9 ~C_8^9. For example, the notation "-(0,1,2,6)" equals to the notation "(3,4,5,7,8)", which means the feature setting is composed of disconnected relation (ID: 3), exclusive argument (ID: 4), exclusive relation (ID: 5), missing argument (ID: 7) and missing relation (ID: 8).

The model using all nine phenomena achieves the best accuracy of 97.62%. Examining the combination sets, we find phenomena IDs 3, 4, 5, 7 and 8 appear quite often in the top 4 feature settings of each combination set. In fact, this setting achieves an accuracy of 95.24%, which is the best performance in C_5^9 combination set. On the one hand, adding more phenomena into (3, 4, 5, 7, 8) setting does not have much performance difference. On the other hand, removing some phenomena from (3, 4, 5, 7, 8) setting or adopting features rather than these phenomena decreases the performance. The best performance of using the feature setting (-(0,6)), i.e., only 7 phenomena, is the same as that of using all 9 phenomena shown in Table 2.

Table 4. Accuracy of combination of negative entailment phenomena.

C_8^9		C_7^9		C_6^9		C_5^9	
-(6)	97.62%	-(0,6)	97.62%	-(0,1,6)	96.67%	-(0, 1,2,6)	95.24%
-(0)	97.62%	-(0,1)	97.14%	-(0,2,6)	96.19%	-(0,1,3,6)	94.29%
-(1)	97.14%	-(1,6)	96.67%	-(0,1,2)	96.19%	-(1,2,3,6)	93.33%
-(2)	96.67%	-(2,6)	96.67%	-(1,2,6)	95.71%	-(0,2,3,6)	93.33%
C_4^9		C_3^9		C_2^9		C_1^9	
(4,5,7,8)	92.38%	(4,7,8)	88.57%	(4,7)	79.52%	(7)	69.52%
(3,4,7,8)	91.43%	(3,4,7)	85.24%	(7,8)	79.05%	(8)	68.57%
(2,4,7,8)	90.48%	(0,7,8)	84.76%	(4,8)	78.57%	(4)	61.90%
(3,4,5,7)	90.00%	(4,5,7)	84.29%	(0,8)	76.67%	(0)	60.95%

We follow Sammons *et al.*'s definitions (2010) and describe the five significant negative entailment phenomena (3, 4, 5, 7, 8) as follows.

(a) **Disconnected Relation**: The arguments and the relations in H are all matched by counterparts in T. None of the arguments in T is connected to the matching relation.

(b) **Exclusive Argument**: There is a relation common to both H and T, but one argument is matched in a way that makes H contradict T.

(c) **Exclusive Relation**: There are two or more arguments in H that are also related in T, but by a relation that means H contradicting T.

(d) **Missing Argument**: Entailment fails because an argument in H is not present in T, either explicitly or implicitly.

(e) **Missing Relation**: Entailment fails because a relation in H is not present in T, either explicitly or implicitly.

The correlations between these five phenomena are shown in Table 5. Each row presents the T-H pairs which are labelled with the corresponding negative entailment phenomenon by the scheme "1 OR 2". Each column in each row denotes the percentage of the T-H pairs which are also labelled with another negative entailment phenomenon. For example, the number of the T-H pairs which are labelled with "Disconnected Relation" is 14, and 2 of the 14 T-H pairs are also labelled with "Missing Argument". Therefore, the column "Missing Argument" in the row "Disconnected Relation" shows the number 2/14 = 14.29%. Table 5 shows the low correlations between most significant negative entailment phenomena. In other words, these phenomena are complementary.

Table 5. Correlations between the five significant negative entailment phenomena.

	Disconnected Relation	Exclusive Argument	Exclusive Relation	Missing Argument	Missing Relation
Disconnected Relation	100.00%	0.00%	0.00%	14.29%	42.86%
Exclusive Argument	0.00%	100.00%	8.70%	8.70%	8.70%
Exclusive Relation	0.00%	16.67%	100.00%	0.00%	16.67%
Missing Argument	4.88%	4.88%	0.00%	100.00%	41.46%
Missing Relation	15.38%	5.13%	5.13%	43.59%	100.00%
Number of Occurrences	14	23	12	41	39

In the above experiments, we do all the analyses on the corpus annotated with linguistic phenomena by human. In some sense, we aim at knowing the ultimate performance of TE recognition systems embodying human knowledge in the inference. Of course, the human knowledge in the inference cannot be captured by TE recognition systems fully correctly. In the later experiments, we explore the five critical features, (3,4,5,7,8), and examine how the performance is achieved if they are extracted automatically.

3. Negative Entailment Phenomena Extraction

The experimental results in Section 2.2 show that disconnected relation, exclusive argument, exclusive relation, missing argument, and missing relation are significant. Our experiments show the combination of these five phenomena is even more powerful. Vanderwende *et al.* (2006) suggested some phenomena that are the clue to false entailments. To model the annotator's inference process, we must first determine the arguments and the relations existing in T and H, and then align the arguments and relations in H to the related ones in T. It is easy for human to find the important parts in a text description in the inference process, but it is challenging for a machine to determine what words are important and what are not, and to detect the boundary of arguments and relations. Moreover, two arguments (relations) of strong semantic relatedness is not always literal identical.

In the following, two methods are proposed to extract the phenomena from T-H pairs automatically in Section 3.2 and Section 3.3. The pre-processing of the pairs is described in Section 3.1.

3.1 Preprocessing

Before extraction, the English T-H pairs are pre-processed according to following considerations.

(a) **Numerical Character Transformation**: All the numerical values are normalized to a single format. The fractional numbers and percentages are converted to real numbers.

(b) **Stemming**: The stemming is performed to each word in the T-H pair with NLTK (Bird, 2002).

(c) **Part-of-Speech Tagging**: Stanford Parser is performed to tagging each word in the T-H pair (Levy & Manning, 2003).

(d) **Dependency Parsing**: Stanford Parser also generates the dependency pairs from T and H (de Marneffe *et al*., 2006). The results of dependency parsing contain crucial information for capturing negative entailment phenomena.

3.2 A Rule-Based Method

Noun phrases are the fundamental elements for comparing the existences of entailment. Given a T-H pair, we first extract 4 sets of noun phrases based on their POS tags: {noun in H}, {named entity (nnp) in H}, {compound noun (cnn) in T}, and {compound noun (cnn) in H}. Then, we extract 2 sets of relations: {relation in H} and {relation in T}, where each relation in the sets is in a form of *Predicate*(*Argument1*, *Argument2*). Some typical examples of relations are *verb*(*subject*, *object*) for verb phrases, *neg*(*A*, *B*) for negations, *num*(*Noun*, *number*) for numeric modifier, and *tmod*(*C*, *temporal argument*) for temporal modifier. A predicate has only 2 arguments in this representation. Thus, a di-transitive verb is in terms of two relations.

Instead of measuring the relatedness of T-H pairs by comparing T and H on the predicate-argument structure (Wang & Zhang, 2009), our method tries to find the five negative entailment phenomena based on the similar representation. Each of the five negative entailment phenomena is extracted as follows according to their definitions. To reduce the error propagation which may be arisen from the parsing errors, we directly match those nouns and named entities appearing in H to the text in T. Furthermore, we introduce WordNet to align synonyms in H and T.

(a) **Disconnected Relation**: If (1) for each $a \in$ {noun in H}$\cup${nnp in H}$\cup${cnn in H}, we can find $a \in$ T too, and (2) for each $r_1=h(a_1,a_2) \in$ {relation in H}, we can find a relation $r_2=h(a_3,a_4) \in$ {relation in T} with the same header h, but with different arguments, i.e., $a_3 \neq a_1$ and $a_4 \neq a_2$, then we say the T-H pair has the "Disconnected Relation" phenomenon.

(b) **Exclusive Argument**: If there exist a relation $r_1=h(a_1,a_2) \in${relation in H}, and a relation $r_2=h(a_3,a_4) \in${relation in T} where both relations have the same header h, but either the pair (a_1,a_3) or the pair (a_2,a_4) is an antonym by looking up WordNet, then we say the T-H pair has the "Exclusive Argument" phenomenon.

(c) **Exclusive Relation**: If there exist a relation $r_1=h_1(a_1,a_2) \in${relation in T}, and a relation $r_2=h_2(a_1,a_2) \in${relation in H} where both relations have the same arguments, but h_1 and h_2 have the opposite meanings by consulting WordNet, then we say that the T-H pair has the "Exclusive Relation" phenomenon.

(d) **Missing Argument**: For each argument $a_1 \in${noun in H}$\cup${nnp in H}$\cup${cnn in H}, if there does not exist an argument $a_2 \in$T such that $a_1=a_2$, then we say that the T-H pair has "Missing Argument" phenomenon.

(e) **Missing Relation**: For each relation $r_1=h_1(a_1,a_2) \in${relation in H}, if there does not exist a relation $r_2=h_2(a_3,a_4) \in${relation in T} such that $h_1=h_2$, then we say that the T-H pair has "Missing Relation" phenomenon.

3.3 A Machine Learning Method

We aim at finding meta-features to describe the characteristic of negative entailment phenomena, and use them for classification. We analyse the dependencies in a T-H pair with Stanford dependency parser (de Marneffe *et al*., 2006) and derive two dependency sets DT and DH for T and H, respectively, where a dependency $gr(g,d)$ is in terms of a binary grammatical relation gr between a governor g and a dependent d. We further define the following three multisets to capture the relationships between T and H:

(a) {H only}=$\{gr | gr(g,d) \in D_H - (D_T \cap D_H)\}$

(b) {Partially identical in governor}=$\{gr | gr(g,d_1) \in D_T,\ gr(g,d_2) \in D_H,\ d_1 \neq d_2\}$

(c) {Partially identical in dependent}=$\{gr | gr(g_1,d) \in D_T,\ gr(g_2,d) \in D_H,\ g_1 \neq g_2\}$

A T-H pair is represented as a feature vector (V(a), V(b), V(c)), where the dimensions of the three vectors V(a), V(b), and V(c) are the number of grammatical relations in the dependency parser. The weights of each grammatical relation *gr* in V(a), V(b), and V(c) are the number of *gr* appearing in the multisets {H only}, {Identical in governor only} and {Identical in dependent only}, respectively. The SVM classifier with the RBF kernel is adopted to develop classifiers with the parameters (cost and gamma) tuned by grid search and evaluated with 10-fold cross validation.

3.4 Experiments and Discussion

The following two datasets are used in English TE recognition experiments.

(a) 210 pairs from part of RTE-5 test set: The 210 T-H pairs are annotated with the linguistic phenomena by human annotators in the work of Mark Sammons *et al* (2010). They are selected from the 600 pairs in RTE-5 test set, including 51% ENTAILMENT and 49% NO ENTAILMENT.

(b) 600 pairs of RTE-5 test set: The original RTE-5 test set, including 50% ENTAILMENT and 50% NO ENTAILMENT.

Table 6 shows the performances of the negative entailment phenomena detection by rule-based and machine-learning methods. The performances of rule-based model are especially poor. The major challenge is to identify the arguments in T-H pairs. (S3) shows an instance. The correct arguments of H in (S3) are “Fifth Amendment right” and “driving license”, but the arguments captured by our method are “Fifth Amendment” and “license”. The issue can be improved with a better dependency parser.

(S3) **T**: “There is a rational basis to distinguish between people driving cars and semi trucks,” Jambois said. “All I would say is I think he has an uphill battle.” The lawsuit says the truckers' **Fifth** and Fourteenth **amendment rights** are being violated because there is no way for them to apply for an occupational license. Mutschler said the state is taking away the truckers' right to drive a truck for a living. He said he will argue that while **driving** is a privilege, once a person has a **license** for work, it becomes a right.

H: **Fifth Amendment right** is about **driving license**.

Table 6. Performance of negative entailment phenomena detection. Reported in Precision (P), Recall (R), and F-Score (F).

Aspect	Rule-based			Learning-based		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
Disconnected Relation	9.52	28.57	14.28	15.91	100.00	27.45
Exclusive Argument	12.94	47.83	20.37	15.49	95.65	26.66
Exclusive Relation	5.71	33.33	9.75	10.43	100.00	18.89
Missing Argument	32.11	37.81	34.72	38.46	97.56	55.17
Missing Relation	23.08	61.54	33.57	32.23	100.00	48.75
Average	16.67	41.82	22.54	22.50	98.64	35.38

Although the rule-based method is poorly-performed, and the machined learning method is not so good at precision and F-Score, the resulting models for TE recognition achieve decent performances. These interesting results are depicted in Table 7. The “Human-annotated” column shows the performance achieved by using the phenomena annotated by human. Using “Human-annotated” phenomena can be seen as the upper-bound of the experiments. In data set (a), the performance of using all the 5 phenomena as features by the machine learning method (M2) is better than that of using the rule-based method (M1). However, the results are reverse in data set (b). This may be because data set (b) contains some cases that cannot be recognized by the model trained from the T-H pairs annotated by human. On the other hand, the rule-based method is implemented directly from the definitions, which is more robust.

Though the performance of using the phenomena extracted automatically by machine is not comparable to that of using the human annotated ones, the accuracy achieved by using only 5 features (59.17%) is just a little lower than the average accuracy of all runs in RTE-5 formal runs (60.36%) (Bentivogli *et al.*, 2009). It shows that the significant phenomena are really effective in dealing entailment recognition even though the phenomena detector is extremely simple. If we can improve the performance of the automatic phenomena detection algorithm, it may make a great progress on the textual entailment.

So far the experiments are two-stage classification. In the first stage, we perform the rule-based or the learning-based model to extract the five negative entailment phenomena. And then, the presences of the five phenomena are used as binary features to recognize the TE in the second stage. In this perspective, the features used for phenomena extraction in Section 3.3 are the *meta-features* of M2. In order to understand the impact of error-propagation, we train a one-stage TE recognizer, M3, by using the meta-features of M2 as features directly. Table 8 compares M1, M2, and M3. The models M2 and M3 do the TE recognition according to the same information, but the two-stage classifier M2 slightly outperforms M3. This result suggests that the concept of negative entailment phenomena is useful for TE recognition.

Table 7. Accuracy of textual entailment recognition using the extracted phenomena as features.

	Dataset (a): 210 pairs			Dataset (b): 600 pairs	
	Rule-based (M1)	Learning-based (M2)	Human-annotated	Rule-based (M1)	Learning-based (M2)
Disconnected Relation	50.95%	54.76%	57.62%	54.17%	51.17%
Exclusive Argument	50.95%	50.95%	61.90%	55.67%	51.83%
Exclusive Relation	50.95%	52.38%	56.67%	51.33%	50.67%
Missing Argument	53.81%	57.62%	69.52%	56.17%	57.33%
Missing Relation	50.95%	50.95%	68.57%	52.83%	55.17%
All	52.38%	60.00%	95.24%	59.17%	57.83%

Table 8. Accuracies of two-stage and one-stage classification.

Stages of Classification	Model	Feature Source	Dataset (a): 210 pairs	Dataset (b): 600 pairs
Two-Stage	M1	Rule-based	52.38%	59.17%
	M2	Machine learning	60.00%	57.83%
One-Stage	M3	Meta-features of M2	56.19%	57.00%

4. Negative Entailment Phenomena in Chinese RITE Dataset

To make sure if negative entailment phenomena exist in other languages, we apply the methodologies in Sections 2 and 3 to the dataset of RITE-1 BC-CT task in NTCIR-9. This dataset contains total 900 traditional Chinese T-H pairs, including 50% ENTAILMENT and 50% NO ENTAILMENT. We annotate all the nine negative entailment phenomena on Chinese T-H pairs according to the definitions by Sammons *et al* (2010) and analyze the effects of various combinations of the phenomena on the new annotated Chinese data. To avoid the influence from the actual entailment label (ENTAILMENT/NO ENTAILMENT), annotators can only see the part of T and H.

Table 9 shows the performances of TE recognition in Chinese with the human knowledge. The interpretation of this table is the same as that of Table 4. The accuracy of using all the nine phenomena as features (i.e., C_9^9 setting) is 91.11%. It shows the same tendency as the analyses on English data. The significant negative entailment phenomena on Chinese data, i.e., (3,4,5,7,8), are the same as those on English data. Besides, we can use only six phenomena to achieve the same performance as using all nine phenomena as features. Furthermore, we also classify the entailment relation by the phenomena extracted automatically by the rule-based method. The process is similar to those of English text described in Section 3.1 and Section 3.2, while Additional effort of processing is required for Chinese text. We segmented Chinese words with Stanford word segmenter (Chang *et al*., 2008) and performed Chinese dependency parsing using Stanford parser and the CNP parser (Chen *et al*., 2009). We extract two sets of negative entailment phenomena according to the parsing results of Stanford parser and CNP parser separately. Both sets are used as independent features to achieve a better performance.

Table 9. Accuracy of combination of negative entailment phenomena on Chinese data.

C_8^9		C_7^9		C_6^9		C_5^9	
-(1)	91.11%	-(1,6)	91.11%	-(1,2,6)	91.11%	-(0,1,2,6)	90.78%
-(2)	91.11%	-(1,2)	91.11%	-(0,1,2)	90.78%	-(1,2,3,6)	89.67%
-(6)	91.11%	-(2,6)	91.11%	-(0,1,6)	90.78%	-(1,2,6,8)	89.33%
-(0)	90.78%	-(0,1)	90.78%	-(0,2,6)	90.78%	-(0,2,4,6)	89.22%
C_4^9		C_3^9		C_2^9		C_1^9	
(3,4,5,7)	89.00%	(3,5,7)	86.11%	(3,7)	80.67%	(7)	74.89%
(3,5,7,8)	87.89%	(4,5,7)	84.78%	(5,7)	80.22%	(8)	67.89%
(0,4,5,7)	87.89%	(0,5,7)	84.67%	(4,7)	79.44%	(0)	56.89%
(1,3,5,7)	87.44%	(2,5,7)	83.89%	(0,7)	79.33%	(4)	56.67%

The rule-based method obtains a similar result of TE recognition in Chinese. The accuracy achieved by using the five automatically extracted phenomena as features is 57.11%, and the average accuracy of all runs in NTCIR-9 RITE task is 59.36% (Shima *et al*., 2011). Compared to other methods using a lot of features, only 12 binary features are used in our method.

5. Conclusion

In this paper we conclude that the negative entailment phenomena have a great effect in dealing with TE recognition. The systems with human annotated knowledge achieve very good performance. Experimental results show that not only can it be applied to the English TE problem, but also has the similar effect on the Chinese TE recognition. To automatically capture the negative entailment phenomena in the text, we propose the phenomenon extraction algorithms with the rule-based and the learning-based approaches. Though the automatic extraction of the negative entailment phenomena still needs a lot of efforts, it gives us a new direction to deal with the TE problem. The fundamental issues such as determining the boundary of the arguments and the relations, finding the implicit arguments and relations, verifying the antonyms of argument and relations, and determining their alignments need to be further examined to extract correct negative entailment phenomena. Besides, multi-class TE recognition will be explored in the future.

Reference

Androutsopoulos, I. & Malakasiotis, P. (2010). A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research, 38*, 135-187.

Bentivogli, L., Clark, P., Dagan, I., Dang, H. T., & Giampiccolo, D. (2011). The seventh PASCAL recognizing textual entailment challenge. In *Proceedings of the 2011 Text Analysis Conference* (*TAC 2011*), Gaithersburg, Maryland, USA.

Bentivogli, L., Dagan, I., Dang, H. T., Giampiccolo, D., & Magnini, B. (2009). The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the 2009 Text Analysis Conference* (*TAC 2009*), Gaithersburg, Maryland, USA.

Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, 69-72.

Chang, P.-C., Galley, M., & Manning, C. (2008). Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proceedings of the Third Workshop on Statistical Machine Translation* (*StatMT '08*), 224-232.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: a Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1-27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, W., Kazama, J., Uchimoto, K., & Torisawa, K. (2009). Improving Dependency Parsing with Subtrees from Auto-Parsed Data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (*EMNLP2009*), 570-579, Singapore.

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. *Lecture Notes in Computer Science*, 3944:177-190.

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (*LREC 2006*), 449-454.

Iftene, A. & Moruz, M. A. (2009). UAIC Participation at RTE5. In *Proceedings of the 2009 Text Analysis Conference* (*TAC 2009*), Gaithersburg, Maryland, USA.

Levy, R. & Manning, C. D. (2003). Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (*ACL 2003*), 439-446.

Sammons, M., Vydiswaran, V.G.V., & Roth, D. (2010). Ask not what textual entailment can do for you... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (*ACL 2010*), Uppsala, Sweden, 1199-1208.

Shima, H., Kanayama, H., Lee, C.-W., Lin, C.-J., Mitamura, T., Miyao, Y. *et al.* (2011). Overview of NTCIR-9 RITE: Recognizing inference in text. In *Proceedings of the NTCIR-9 Workshop Meeting*, Tokyo, Japan.

Vanderwende, L., Menezes, A., & Snow, R. (2006). Microsoft Research at RTE-2: Syntactic Contributions in the Entailment Task: an implementation. In *Proceedings of the Second PASCAL Challenges Workshop.*

Wang, R. & Zhang, Y. (2009). Recognizing Textual Relatedness with Predicate-Argument Structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 784–792.

Watanabe, Y., Miyao, Y., Mizuno, J., Shibata, T., Kanayama, H., Lee, C.-W. *et al*. (2013). Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. In *Proceedings of the 10th NTCIR Conference*, 385-404, Tokyo, Japan

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member： NT$ 1,000 (US$50.-)
group membership： NT$20,000 (US$1,000.-)
life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#： ________________

Name： __________________________________ Date of Birth： _______________

Country of Residence： ____________________ Province/State： ____________________

Passport No.： ___________________________ Sex: _______________________

Education(highest degree obtained)： ____

Work Experience： ____

____

Present Occupation： ____

Address： ____

____

Email Add： ____

Tel. No： ____________________________ Fax No： ___________________________

Membership Category：□ Regular Member □ Life Member

Date： _____/____/____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member ： US$ 50.- （NT$ 1,000）
Life Member ： US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）

（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目

（三）收集國內外有關計算語言學知識之圖書及最新發展之資料

（四）發行有關之學術刊物，論文集及通訊

（五）研定有關計算語言學專用名稱術語及符號

（六）與國際計算語言學學術機構聯繫交流

（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

會員類別	□終身 □個人 □學生	會員編號		（由本會填寫）	
姓　　名		性別		出生日期	年　　月　　日
				身分證號碼	
現　　職		學　　歷			
通訊地址	□□□				
戶籍地址	□□□				
電　　話		E-Mail			
申請人:　　　　　　　　（簽章） 中　華　民　國　　　年　　　月　　　日					

審查結果:

1. 年費：

   終身會員：　10,000.-
   個人會員：　1,000.-
   學生會員：　500.-（限國內學生）
   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638
   E-mail：aclclp@hp.iis.sinica.edu.tw　　網址: http://www.aclclp.org.tw
   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)
PAYMENT FORM

Name: ___________________________________(Please print) Date: _______________

Please debit my credit card as follows: US$ _____________

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD Issue Bank:__________________

Card No.: ________ -_________-________ -_________ Exp. Date:____________(M/Y)

3-digit code: _________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE: ___________________________________

Phone No.: _________________________E-mail: ___________________________________

Address: ____

PAYMENT FOR

US$ _____ ❑ Computational Linguistics & Chinese Languages Processing (IJCLCLP)

Quantity Wanted: ______

US$ _____ ❑ Journal of Information Science and Engineering (JISE)

Quantity Wanted: ______

US$ _____ ❑ Publications:________________________________

US$ _____ ❑ Text Corpora: ______________________________

US$ _____ ❑ Speech Corpora:_____________________________

US$ _____ ❑ Others: __________________________________

US$ _________ ❑ Membership Fees ❑ Life Membership ❑ New Membership ❑Renew

US$ _________ = Total

Fax 886-2-2788-1638 or Mail this form to:

ACLCLP
℅ IIS, Academia Sinica
Rm502, No.128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中華民國計算語言學學會
信用卡付款單

姓名: ____________________________(請以正楷書寫)　日期:：____________________

卡別：❑ VISA CARD　❑ MASTER CARD ❑ JCB CARD　發卡銀行：______________

信用卡號：__________-__________-__________-__________　有效日期：__________(m/y)

卡片後三碼：______________（卡片背面簽名欄上數字後三碼）

持卡人簽名：_______________________________________(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：____________________________E-mail：_________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$__________ ❑ 中文計算語言學期刊(IJCLCLP) _______________

NT$__________ ❑ Journal of Information Science and Engineering (JISE)

NT$__________ ❑ 中研院詞庫小組技術報告_____________________

NT$__________ ❑ 文字語料庫 ____________________________________

NT$__________ ❑ 語音資料庫 ____________________________________

NT$__________ ❑ 光華雜誌語料庫1976~2010

NT$__________ ❑ 中文資訊檢索標竿測試集/文件集

NT$__________ ❑ 會員年費：❑續會　❑新會員　❑終身會員

NT$__________ ❑ 其他: _____________________________________

NT$_______________ = 合計

填妥後請傳真至 02-27881638 或郵寄至:
11529台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

		Surface	AIR (US&EURP)	AIR (ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本)　ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02　V-N 複合名詞討論篇 & 92-03　V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05　中文詞類分析	10	15	13	_____	_____
7.	no.93-06　現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02　古漢語字頻表	11	16	14	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03　訊息爲本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02　論語詞頻表	9	14	12	_____	_____
16.	no.98-01　詞頻詞典	18	30	26	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03　自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01　現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT：□ Credit Card (Preferred)
  □ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息爲本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01　「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義 （中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期）　年份：______ （過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※　此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人：　黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：　________________________　收據抬頭：______________________________

地　　址：　____

電　　話：　__________________________　E-mail: _____________________________

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright：It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.

1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.

2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.

3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.

4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).

5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript

6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.

7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

> Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

> Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).

Please visit the following websites for details.

(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)

(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

Online Submission: http://www.aclclp.org.tw/journal/submit.php

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.19 no.3 September 2014

Contents

Papers

章無疵也章之明靡句
無玷也句之清英字不
妄也文賦曰選義按部
考辭就班就所傳達者
觀之禮記曰發志為言
發言為名傳曰言以足志
文以足言易曰書不盡言
言不盡意詩序曰在心為
志發言為詩情動於中
而形於言蓋情志發而
語言成語言工而文字傳
也

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing