An Inference Problem Set for Evaluating Semantic Theories and Semantic Processing Systems for Japanese

Ai Kawazoe[1], Ribeka Tanaka[2], Koji Mineshima[2], and Daisuke Bekki[2]

[1] National Institute of Informatics, 2-1-2 Hitotsubashi Chiyoda-ku, Tokyo, Japan
[2] Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku, Tokyo, Japan

Abstract. This paper introduces a collection of inference problems intended for use in evaluation of semantic theories and semantic processing systems for Japanese. The problem set categorizes inference problems according to semantic phenomena that they involve, following the general policy of the FraCaS test suite. It consists of multilingual and Japanese subsets, which together cover both universal semantic phenomena and Japanese-specific ones. This paper outlines the design policy used in constructing the problem set and the contents of a beta version, currently available online.

1 Introduction

Explaining the validity (or invalidity) of inference (e.g. entailment, presupposition, and implicature) among sentences is one of the main objectives of studies of meaning. Since the start of the PASCAL RTE challenge (Dagan et al. 2006), the recognition of such inference relations has been a core component of NLP tasks, and the necessity and importance of inference problem sets is widely recognized. Aiming to contribute to the development and evaluation of semantic theories and semantic processing systems for Japanese, we are constructing a data set which comprises inference problems involving Japanese semantic phenomena. For English, the FraCaS test suite (Cooper et al. 1996), which covers major semantic phenomena, has been used for textual entailment (TE) recognition tasks, but no such data set has previously been constructed for Japanese. Our problem set consists of two parts: a multilingual subset and a Japanese subset. The multilingual subset includes Japanese counterparts of FraCaS problems. The Japanese subset covers some universal phenomena not included in FraCaS and some specific to Japanese, such as *toritate* particles and *wa–ga* constructions. In this paper, we outline the design policy used in constructing the problem set and describe a beta version that we have released online.

Each inference problem in the original FraCaS test suite is a triplet: a premise or set of premises; a yes/no question; and the answer to that question. A "hypothesis" sentence, a declarative counterpart of the yes/no question, have since been added in a machine-readable version by Bill MacCartney.[3] In the following sections, we simply omit questions and show only the premises (P), hypotheses (H), and answers when illustrating inference problems, exemplified as follows.

[3] `http://www-nlp.stanford.edu/~wcmac/downloads/fracas.xml`

(1) fracas-141 answer:unknown

P1 John said Bill had hurt himself.

H Someone said John had been hurt.

2 Background

When evaluating linguistic theories, the primary requirement for evaluation data is that each data item represents only the target phenomena for that item to the greatest possible extent. For inference relations, it is ideal that the data involve only the target phenomena, and do not include other factors that could affect speakers' judgments of the validity of an inference between a premise set and a hypothesis. This requirement is also important for data to be used in NLP tasks such as textual entailment. Recently, several researchers have pointed out the necessity of data that can allow measuring system performance on specific phenomena (e.g., Bos 2008; Sammons et al. 2010; Bentivogli et al. 2010; Cabrio et al. 2013).

The FraCaS test suite was created by the FraCaS Consortium as a benchmark by which to measure the semantic competence of NLP systems. It contains 346 inference problems that collectively demonstrate basic linguistic phenomena for which formal semantics should account; these include quantification, plurality, anaphora, ellipsis, tense, comparatives, and propositional attitudes. Each problem is designed to include exactly one target phenomenon, to exclude other phenomena, and to be independent of background knowledge. A machine-readable version of FraCaS has been used for evaluation and error analyses of several TE models (e.g., MacCartney and Manning 2007, 2008; Lewis and Steedman 2013; Tian et al. 2014; Abzianidze 2015; Mineshima et al. 2015). Currently, the MultiFraCaS project, headed by Robin Cooper, is working to create a multilingual FraCaS test suite.[4]

One strength of FraCaS-type data sets is that they are based on the outcomes of standard linguistic studies and thus represent reliable observations of phenomena. This means that the quality of the data set (e.g., the accuracy of judgements about the validity of inference, and the validity of analyses) is ensured by the community of linguists. In addition, FraCaS-type data sets have enough generality that the validity of inference will usually not be changed if we replace content words in the sentences and the situation of utterance. This is because each problem represents a generalization about relevant semantic phenomena.

For the Japanese language, some existing inference problem sets have been designed so as to restrict the number of phenomena or factors that affect possibilities of inference. For example, NTCIR RITE provides the UnitTest data set, following the methodologies in Sammons et al. (2010) and Bentivogli et al. (2010). The Kyodai Textual Entailment Evaluation Data, which was created by Kotani et al. (2008), contains premise–hypothesis pairs with only one or two factors relevant to the validity of inference. However, many of the problems in these sets involve lexical, syntactic, or world knowledge; further, basic semantic phenomena, such as those in FraCaS, are not fully represented. A set with a wider variety of inference problems should be created to cover major semantic phenomena of Japanese.

[4] http://www.ling.gu.se/~cooper/multifracas/

3 Overview of the problem set

3.1 Multilingual and Japanese Subsets

Our problem set categorizes inference problems according to the semantic phenomena they involve, following the design policy of the FraCaS test suite. The content of the problem set is shown in Table 1. In that table, phenomena marked with "⋆" are those covered by the beta version.

Subsets	Descriptions	Sections	Num.
Multilingual subset	Japanese counterparts of FraCaS Problems	⋆Generalized quantifier, ⋆plurality, ⋆anaphora, ⋆ellipsis, ⋆adjectives, ⋆comparatives, ⋆tense, ⋆verbs, ⋆propositional attitude	624
Japanese subset	Problems with universal phenomena not covered by FraCaS	Modality, conditionals, negation, ⋆adverbs, focus, and more phenomena with ⋆adjectives, ⋆verbs, ⋆comparatives and ⋆propositional attitudes, etc.	166
	Problems with Japanese-specific phenomena	⋆*Toritate* particles, *wa–ga* construction, etc.	

Table 1. Sections of our problem set. Those covered by the beta version as of April 2015 are denoted by "⋆".

The multilingual subset of our problem set contains Japanese counterparts of FraCaS problems, but we have not adhered to the principle of one-to-one correspondence that is followed for the MultiFraCaS test sets. As a result, 90 of the FraCaS problems correspond to more than one of the problems in our data set. In particular, those FraCaS problems with generalized quantifiers have many Japanese counterparts, which was done because there are many Japanese expressions and word patterns that are truth-conditionally equivalent to quantificational NPs in English (but may introduce different presuppositions and/or implicatures).

Not all problems in the multilingual subset are literal translations of FraCaS problems. For those FraCaS items that have no natural translation, we created Japanese problems that target similar phenomena, albeit with different syntactic structures or lexical items. The majority of the Japanese subset is still being developed. We have covered more phenomena involved with adjectives and verbs which are not covered by FraCaS and some phenomena with adverbs and toritate particles.

3.2 Format

We adopted the following format for our problem set.

- `problem`: an inference test
  - `jsem_id` attribute: an unique ID

    - `answer` attribute: validity of inference (yes, no, unknown, or undef)
    - `phenomena` attribute: type of phenomena (multiple entries allowed)
    - `inference_type` attribute: type of inference
- `link`: a link to a resource in other languages
    - `resource` attribute: the name of the linked resource
    - `link_id` attribute: the ID of the corresponding test in the linked resource
    - `translation` attribute: specifies whether the inference test is a literal translation of the linked test or not. (allowed: yes, no, unknown)
    - `same_phenomena` attribute: specifies whether the inference test represents the same phenomena as the linked test or not (allowed: yes, no, unknown)
- `p`: premise
- `h`: hypothesis
- `note`: comments

For example, the Japanese equivalent of fracas-141 (shown above as (**??**)) is described as follows.

`problem`	`id`: 449
	`answer`: yes
	`phenomena`: Nominal anaphora, intra-sentential anaphora, zibun
	`inference_type`: entailment
`link`	`resource`: fracas
	`link_id`: 141
	`translation`: yes
	`same_phenomena`: no
`p1`	ジョンは、ビルが自分を傷つけたと言った。
`English`	John said Bill had hurt himself.
`h`	誰かが、ジョンが傷つけられたと言った。
`English`	Someone said John had been hurt.
`note`	As is well-known, unlike “himself,” “zibun” allows long-distance anaphora (Kuno 1978). This makes it possible to interpret “John” as the antecedent of “zibun” in p1, which is in contrast to the English counterpart.

Some elements—such as `problem`, `p`, `h`, `note`, and the attribute `answer`—are based on the FraCaS and MultiFraCaS representation. We added some new elements and attributes, as described below.

Links to the FraCaS problems The `link` element is added to show information about a linked resource. The attributes `translation` and `same_phenomena` are introduced to specify translation- and phenomena-level similarities and differences between multi-language problem pairs. As shown in the example above, we can see that the Japanese problem is a possible translation of an English counterpart by looking at the value of the `translation` attribute. The value of the `same_phenomena` attribute shows that they involve disparate phenomena, leading to a different answer (“yes”) for the Japanese version than for the English version (which has answer “unknown”). Details of relevant phenomena and references to relevant literature are given in the `note` element.

Categories of Phenomena Phenomena involved in an inference problem are concisely described by `phenomena` attributes. We allow multiple value entries for this attribute and recommend creating new values as necessary in the process of constructing the problem set. In our construction, created values have been collected afterwards and edited.

Currently the values for `phenomena` in our problem set are classified into three types: section titles, universal phenomena, and Japanese-specific phenomena. Section titles are taken from the nine sections of the original FraCaS (Generalized Quantifiers, Plurals, Nominal Anaphora, etc.). Some of the universal phenomena values are also taken from subsections or problem descriptions in FraCaS, and others were newly created by us (e.g., factive/non-factive/counter-factive for propositional attitude problems; some other types of anaphora, such as coreference and bound variable anaphora). Values that indicate Japanese-specific phenomena include several anaphora types (*no* anaphora, *soo su* anaphora), elliptic constructions (stripping with or without case markers), word order patterns of a quantifier and an NP it modifies (pre- or post-nominal quantifiers, floating quantifiers), and key functional words (anaphoric expressions such as *zibun*, *kare/kanozyo* and *so*-series demonstratives, various conjunctive particles, etc.).

Inference Types Inference is a complex phenomenon that typically involves various linguistic and contextual factors when we judge the validity or invalidity of an inference relation among sentences. In our problem set, we specify the type of inference for each premise–hypothesis pair, using the `inference_type` attribute for this purpose, in addition to specifying the type of phenomenon via the `phenomena` attribute. This enables evaluation of TE models according to inference type.

The major values for the `inference_type` attribute are `entailment` and `presupposition`. Distinction between the two inference types is based on well-known inference classifications in formal semantics and pragmatics (e.g., Chierchia & McConnell-Ginet 2000; Levinson 2000; Kadmon 2001; Potts 2005). Entailment is an at-issue content of utterance, also called "asserted content," "What is Said," (Grice 1989) or "semantic entailment," as distinguished from "entailment" in the broader sense. The problem (**??**), one of the counterparts of fracas-017, shows a typical example of entailment.

(2) id:117 answer:yes
P1 一人のアイルランド人がノーベル文学賞を受賞した。
One-CL-GEN Irishman-NOM Nobel literature prize-ACC win-PAST
"An Irishman won the Nobel Prize in Literature."

H 一人のアイルランド人がノーベル賞を受賞した。
One-CL-GEN Irishman-NOM Nobel prize-ACC win-PAST
"An Irishman won a Nobel Prize."

Presupposition, in contrast, acts as background content for an utterance, and is often indicated by specific lexical items or expressions. The problem (**??**) is an example of presupposition, signified by the factive predicate *koto-o uresiku omou* (lit., be pleased to know that).

(3) id:737 answer:yes

P1 太郎は花子が高校を卒業したことを嬉しく思った。
Taroo-TOP Hanako-NOM high school-ACC graduate-PAST-COMP-ACC pleased think-PAST
"Taro was pleased to know that Hanako graduated from high school."

H 花子は高校を卒業した。
Hanako-TOP high school-ACC graduate-PAST
"Hanako graduated from high school."

Neither entailment nor presupposition can be cancelled by subsequent contexts, but the former disappears and the latter survives when the premise appears in a modal or negated context and when it appears in the antecedent of a conditional.

Conventional and conversational implicature are two other major types of inference, and are important data for TE recognition tasks. Although the current beta version of our problem set covers only problems with entailment and presupposition, we plan to expand it to include these implicature cases.

3.3 Creation of the problem set

Four linguists constructed the problem set. In principle, one linguist constructed inference problems for each section and another reviewed them, revising as necessary. We strongly recommended referring to the relevant literature when introducing new problems. In the review process, we checked for the presence or absence of factors other than targeted factors, for ambiguity and naturalness of sentences, and for cross-rater reliability of inference judgments.

4 Concluding Remarks

We have introduced a FraCaS-type inference problem set that covers semantic phenomena in Japanese. We have shown some new features that may contribute to a cross-linguistic evaluation of TE models according to phenomenon or inference type. The problem set is now being expanded to cover more phenomena, both universal and Japanese-specific ones. We encourage linguists to become collaborators for the data set by contributing inference problems with their specialized knowledge and findings.

References

1. Abzianidze, L. 2015. Towards a Wide-coverage Tableau Method for Natural Logic. 2015. In Murata, T., Mineshima, K., and Bekki, D. (Eds.), New Frontiers in Artificial Intelligence: JSAI-isAI 2014 Workshops, LENLS, JURISIN, and GABA, Revised Selected Papers. Lecture Notes in Computer Science, volume 9067, 66-82.
2. Bentivogli, L., Cabrio, E., Dagan, I., Giampiccolo,D., Leggio, M. L., and Magnini, B. 2010. "Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference." Proceedings of LREC 2010:3544-3549, Valletta, Malta.
3. Bos, J. 2008. "Let's not argue about semantics." Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008):2835-2840, Marrakech, Morocco.
4. Chierchia, G and McConnell-Ginet, S. 2000. Meaning and Grammar: An Introduction to Semantics. MIT Press.

5. Cooper, R., Crouch, D., van Eijck, J., Fox, C., van Genabith, J., Jan, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., Pulman, S., Briscoe, T., Maier, H., and Konrad, K. 1996. "Using the framework." Technical report, FraCaS: A Framework for Computational Semantics. FraCaS deliverable D16.
6. Cabrio, E., and Magnini, B. 2013. "Decomposing Semantic Inferences." C. Condoravi, A. Zaenen (eds.) Linguistic Issues in Language Technology (LiLT), 9(1). Special Issue on The Semantics of Entailment.
7. Dagan, I., Glickman, O., and Magnini, B. 2006. "The pascal recognising textual entailment challenge." Lecture Notes in Computer Science, volume 3944, 177-190.
8. Grice, P. 1989. Studies in the way of words. Harvard University Press.
9. Kadmon, N. 2001. Formal Pragmatics. Blackwell.
10. Kotani, M., Shibata, T., Nakata, T., and Kurohashi, S. 2008. "Building Textual Entailment Japanese Data Sets and Recognizing Reasoning Relations Based on Synonymy Acquired Automatically." Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing, Tokyo, Japan.
11. Kuno, S. 1978. Danwa no bunpoo [grammar of discourse]. Tokyo: Taishukan.
12. Levinson, S. C. 2000. Presumptive Meanings: The theory of generalized conversational Implicature. MIT Press.
13. Lewis M., and Steedman, M. 2013. "Combining distributional and logical semantics." Transactions of the Association for Computational Linguistics, 1:179-192.
14. MacCartney, B., and Manning, C. D. 2007. "Natural logic for textual inference." In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, 193-200.
15. MacCartney, B., and Manning, C. 2008. "Modeling semantic containment and exclusion in natural language inference." The 22nd International Conference on Computational Linguistics (Coling-08), Manchester, UK, August.
16. Mineshima, K., Martínez-Gómez, P., Miyao, Y., and Bekki, D. 2015. Higher-order logical inference with compositional semantics. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), Lisbon, Portugal. 2055-2061.
17. Potts, C. 2005. The Logic of Conventional Implicatures. Oxford University Press. Sammons2010 Sammons, M., Vinod Vydiswaran, V. G., and Roth, D. 2010. "Ask not what textual entailment can do for you..." Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics:1199-1208, Uppsala, Sweden.
18. Tian, R., Miyao, Y., and Matsuzaki, T. 2014. "Logical inference on dependency-based compositional semantics." In Proceedings of ACL, 79-89.