

Software Demonstration: The Sharp Intelligent Dictionary

Pete WHITELOCK and Philip EDMONDS, Oxford, UK

Abstract

This paper describes the Sharp Intelligent Dictionary (SID), an English-Japanese glossing system for Japanese readers and learners of English. SID uses a variety of lightweight analysis techniques, a large bilingual dictionary and a prioritised model of collocations to present informed guesses about the best translations of words and expressions in their context.

1 Introduction

The Sharp Intelligent Dictionary (SID) is a tool to assist Japanese learners to read English texts online by providing interlinear Japanese glosses for the English words and expressions. It aims to help users improve their vocabulary by avoiding the laborious consultation of bilingual paper dictionaries, a characteristic feature of the Japanese learning experience but one which may 'hinder and interrupt the L2 reading comprehension process' [Lomicka, 1998]. SID is available for download as part of the Power E/J package (trial version) from http://www.sharp.co.jp/sc/excite/soft_map/ej5/p_ej50.htm. The figure below shows SID in use:

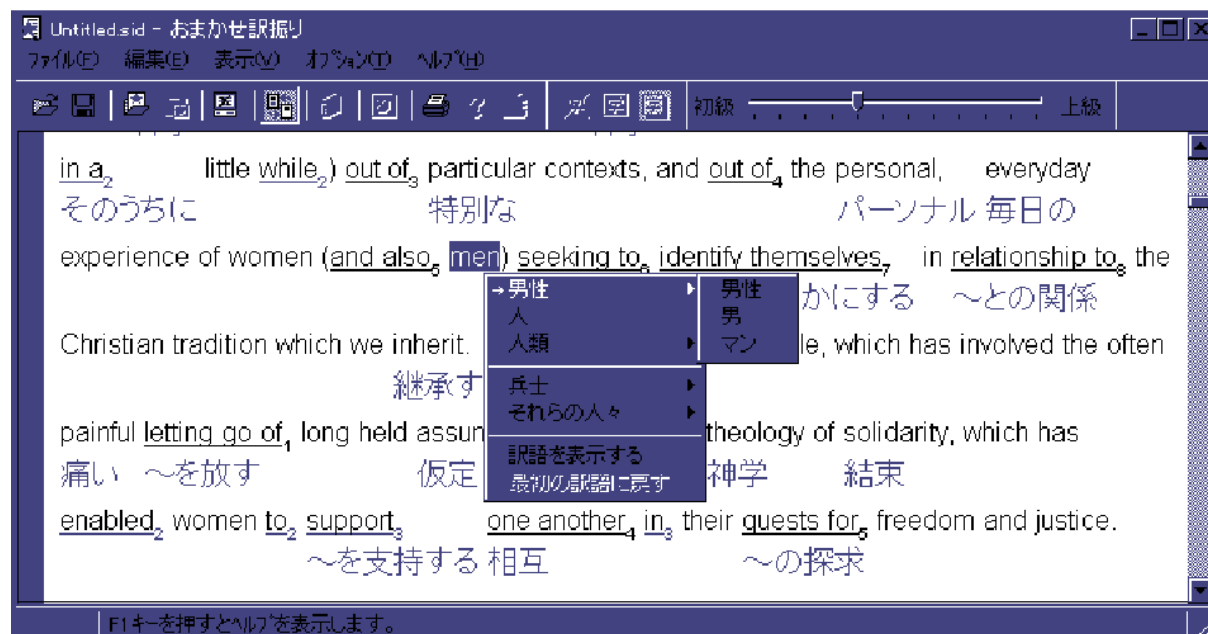


Figure 1: SID in use

SID's innovation over the typical electronic dictionary is its attempt to provide the user with the correct equivalent for a word as used in the passage at hand. This behaviour, which we call contextivity, is based on two technologies. First, a part-of-speech tagger resolves ambiguities in the syntactic categories of words (including transitivity type for verbs, etc.) using a probabilistic model of category sequences. Secondly, a 'collocator' detects multiple words (and smaller objects) that may require glossing as a unit; it then chooses the best consistent set of collocations and displays their glosses. Unlike some similar systems, SID can also detect discontinuous collocations such as the phrasal verbs ubiquitous in idiomatic English, but typically under-represented in the learner's vocabulary.

Of course, SID's guesses at the best Japanese gloss are not perfect. However, the user can choose an alternative from a menu, and the best consistent set of glosses will be adjusted accordingly. Sense ambiguity that is not resolved by syntactic tagging, collocated items, or subject field is treated on a most-frequent-first basis, but the user can choose a gloss for an alternative sense of the same expression and have the system prefer this gloss for the rest of the document, or permanently.

To assist the learner with systematic context-based vocabulary building, SID dictionary entries are divided into a number of difficulty levels according to their frequency, transparency and other factors influencing the likelihood and utility of learning them. The system will display glosses for only those English words that are more difficult than a user-defined level. In this way the information presented can be adjusted to meet the changing needs of the user as reader or vocabulary learner.

2 Background

The Sharp Intelligent Dictionary grew out of research in Shake-and-Bake translation [Whitelock, 1992], an attempt to build a linguistically satisfactory and operationally adequate theory of translation with the bilingual dictionary, in its simplest sense, at the center. SID incorporates a large English-Japanese dictionary developed by our Japanese colleagues in Sharp Corporation over a period of 15 years. We have replaced a deep analysis of the English text with techniques from the corpus-based, statistical paradigm (see [Manning & Schütze, 1999] for an excellent introduction) to establish a shallow but robust analysis of the input text.

3 SID's English analysis

The English analysis component of SID is composed of two phases, which we call correlation and collocation. The correlator establishes a probability distribution for the potential parts-of-speech and corresponding morphemic structures of each input word, and the collocator detects sets of morphemes that will be translated as a unit.

The correlator comprises a finite-state morphological analyser, a Hidden Markov Model bigram tagger and a set of correlation rules. The major components, although developed at Sharp Labs., are essentially off-the-shelf computational models. The data model for the part-of-speech tagger, however, has been enriched using the Penn TreeBank parsed Brown corpus [Marcus et al.

1993], and is now used to make much finer distinctions than those made by typical taggers. The latter tend to keep the tagset coarse-grained in order to achieve high precision. However, if any component of the system needs, say, transitivity information, it is better that the tagger makes an educated guess than reserves all judgement. Furthermore, finer-grained tags can even improve tagger-internal accuracy by providing more sharply delineated contexts on which to train the tagger. For instance, by distinguishing articles (*a, the, no, every*) from other determiners (*this, these, etc.*, which are also pronouns), we give the tagger no reason to believe that *flies* in *the flies* is anything other than a noun – in the Penn tagset its tag sequence would be the same as that of *this flies*.

The SID tagger thus makes a variety of distinctions that are learnable in the bigram model, including transitivity types of verbs, the attributive/predicative status of adjectives and verbs, the head/modifier status of nouns, prepositional/infinitival *to*, subject/object pronouns, and determiner/pronoun status of the relevant items. The auxiliary verbs (*be, have, do*) are distinguished from each other and from main verbs (including main verb *have* and *do*). Already this relatively rich set of tags does much to guide the system to the correct translation. For instance, [Wilks & Stevenson 1998] claim that assignment of even coarse-grained tags serves to distinguish between homographs in 87.4% of ambiguous word tokens. Our richer tagset allows us to go beyond this to discriminate, for example, attributive *present* (i.e., not *past*) from predicative *present* (i.e., not *absent*).

The collocator locates sets of words, or more strictly morphemes, that will be translated together. Consistent with Sinclair's Principle of Idiom [Sinclair 1991], which underscores the primacy of 'semi-preconstructed' phrases, SID typically interprets between 20% and 50% of the word tokens in a text as belonging to a multi-word expression. Such expressions include prepositional and phrasal verbs, compound nouns, collocations with light verbs, formulae such as proverbs and similes, and combinations of open class items with governed morphemes such as the infinitival marker and participial affixes. SID's general language dictionary includes over 25,000 distinct expressions, with the specialised subject lexicons contributing a further 80,000. While not perfect, the detection of such expressions is invaluable in guiding the user to the correct translation. It relieves the burden both of detecting the presence of a (possibly discontinuous) collocation and of locating the relevant section of the dictionary entry – often far from easy in a printed dictionary (see e.g., [Bogaards 1996]).

Although SID does not perform a deep analysis in order to establish that the items of a potential collocation stand in the correct structural relationship, it uses a prioritised tiling scheme [Poznanski et al. 1998] that allows a close approximation to richer schemes of syntactic analysis.

4 Selection of best glosses

The collocator detects all potential collocations and associates with each a priority score, from which is determined the best consistent set for display. This set, which we call the fringe, may use each morpheme in a sentence a maximum of one time; that is, morphemes are resources that are consumed in priority order of the entries that contain them. Items not appearing in the fringe appear in a right-click menu under any of their component words. If the user chooses a collocation that is not on the fringe then this is given maximum priority and the fringe is recomputed by a consistency maintenance system.

The priority score of a collocation is a function of the number of words it contains, their separation, and the probabilities, as assigned by the tagger, that the elements of the sentence conform to the part-of-speech constraints of the collocation. In earlier versions of the system, these factors were used in order: collocations with more words were always preferred to those with fewer words, even if widely separated, and compact collocations were always preferred to less compact ones, even if their tag probabilities were much lower. Thresholds on separation of elements (5 words) and tag probabilities (4%) prevented the (usually incorrect) detection of collocations involving very widely separated items or very unlikely readings. As reported in [Poznanski et al. 1998], the results of this scheme were evaluated at a precision of 82% correct collocation – not sense – detection .

In the latest version of SID, we combined the priority factors into a single integrated metric. We determined, from the parsed Brown corpus and other sources, a table giving the probability that the elements of each class of collocation appear at a certain separation. Under this scheme, we re-evaluated the precision of collocation detection at around 89%. Furthermore, we were able to set more generous thresholds (separation 9 words, tag probability 1%), improving overall recall by licensing unlikely collocations to appear in the menus, without adversely affecting fringe precision.

5 Graded Vocabulary

In the latest version of SID, we have assigned to each word and phrase of English one of 12 difficulty levels. Monolingual learner's dictionaries (LDOCE III, COBUILD II) have recently started to include frequency information, though the number of levels is much smaller. We have taken advantage of the uniformity of the Japanese experience of learning English to define levels corresponding roughly to the year of study. Starting with the vocabulary lists employed by publishers such as Bun-Eido and Taishukan, we updated and refined the levels with frequency information from the British National Corpus, taking steps to remedy the British English bias of the latter.

In a printed learner's dictionary, frequency information serves a single purpose, essentially that of indicating the importance of the word. In SID, the showing or hiding of words according to difficulty level also serves as an instant indicator of the difficulty of the document. Most importantly, it allows the learner to set the display appropriate to her year of study, leaving hidden the glosses of those words she is currently trying to learn; in this way the system encourages the learner to infer the meanings of unknown words in context.

The value to the learner of contextual inference is obvious, though it is neither a reliable skill [Pressley et al. 1987] nor of primary importance in vocabulary acquisition for every learner [Hulstijn 1993]. Its main limitation is that the student is restricted to the study of carefully selected texts in which the percentage of unknown vocabulary is at a sufficiently low level for the guessing of meanings to be a viable strategy. [Laufer 1989] suggests 95% of the words should be known for reasonable comprehension.

SID's graded vocabulary allows the 5% unknown words to be those at a level of difficulty appropriate to the learner, not merely the rarest or most difficult 5%, since the more difficult words will be glossed and thus, in a sense, known. It thus hugely expands the range of texts

available to the learner as pedagogic aids. Incidental vocabulary acquisition is largely a matter of the rich get richer [Horst et al. 1999] – SID offers the possibility that a bilingual gloss – immediate and comprehensible – can stand in for some of these riches, and thus allow less advanced learners to partake in the incidental learning benefits of extensive reading.

6 Prognosis

A bilingual dictionary such as SID addresses the learner's immediate comprehension needs and may facilitate incidental vocabulary acquisition. For truly effective vocabulary building, the learner must have access to more detailed monolingual information about words and phrases in order to explore their properties and contexts, establish their relationships and consolidate the learner's knowledge. SID's contextive technology provides an excellent basis for future research to address this combination of convenience and comprehensive content.

References

- [Bogaards 1996] Paul Bogaards: "Dictionaries for Learners of English", in: *International Journal of Lexicography*, 9 (4), 1996.
- [Horst et al. 1999] Marlise Horst, Tom Cobb and Paul Meara: "Beyond A Clockwork Orange: Acquiring Second Language Vocabulary through Reading", in: *Reading in a Foreign Language*, 11 (2). Also at <http://www.er.uqam.ca/nobel/r21270/cv/Casterbridge.html>
- [Hulstijn 1993] J. Hulstijn: "When do foreign-language readers look up the meaning of unfamiliar words? The influence of task and learner variables", in: *Modern Language Journal* 77.
- [Laufer 1989] B. Laufer: "What percentage of text-lexis is essential for comprehension?" In C. Lauren and M. Nordman (eds.), *Special Language: From Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters.
- [Lomicka 1998] Lara Lomicka "To gloss or not to gloss: An investigation of reading comprehension online", in: *Language Learning & Technology*, Vol. 1, No. 2, January 1998. Also at <http://polyglot.cal.msu.edu/llt/vol1num2/article2/default.html>
- [Manning & Schütze 1999] Christopher Manning and Hinrich Schütze: *Foundations of Statistical Natural Language Processing*, MIT, 1999.
- [Marcus et al. 1993] Mitch Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz: "Building a large annotated corpus of English: the Penn TreeBank", in: *Computational Linguistics*, vol. 19, 1993.
- [Poznanski et al. 1998] Victor Poznanski, Pete Whitelock, Jan IJdens & Steffan Corley: "Practical Glossing as Prioritised Tiling", in: *Proceedings of the 17th COLING/36th ACL Montreal*, 1998.
- [Pressley et al. 1987] M. Pressley, J.R. Levin and M.A. McDaniel: "Remembering versus inferring what a word means: Mnemonic and contextual approaches" In M. McKeown and M. Curtis (eds.) *The Nature of Vocabulary Acquisition*, Lawrence Erlbaum, 1987.
- [Sinclair 1991] John Sinclair: *Corpus, Concordance, Collocation*. OUP, 1991.
- [Wilks & Stevenson 1998] Yorick Wilks and Mark Stevenson: "The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation", in: *Journal of Natural Language Engineering* 4(2).

[Whitelock 1992] Pete Whitelock: "Shake-and-Bake Translation", in: *Proceedings of the 14th COLING*, Nantes, 1992.