A Reexamination of MRD-based Word Sense Disambiguation

Timothy Baldwin,♠ Su Nam Kim,♠ Francis Bond,♡ Sanae Fujita,♢ David Martinez♣ and Takaaki Tanaka‡

♠ Department of Computer Science and Software Engineering
University of Melbourne, VIC 3010, Australia

♡ Division of Linguistics and Multilingual Studies
Nanyang Technological University
Level 3, Room 55, 14 Nanyang Drive, Singapore 637332

♢ NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
2-4 Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

♣ NICTA Victoria Research Laboratories
University of Melbourne, VIC 3010, Australia

‡ Research and Development Center,
Nippon Telegraph and Telephone West Corporation
6-2-82 Shimaya Konohana-ku Osaka 554-0024, Japan

This paper reconsiders the task of MRD-based word sense disambiguation, in extending the basic Lesk algorithm to investigate the impact on WSD performance of different tokenisation schemes and methods of definition extension. In experimentation over the Hinoki Sensebank and the Japanese Senseval-2 dictionary task, we demonstrate that sense-sensitive definition extension over hyponyms, hypernyms and synonyms, combined with definition extension and word tokenisation leads to WSD accuracy above both unsupervised and supervised baselines. In doing so, we demonstrate the utility of ontology induction and establish new opportunities for the development of baseline unsupervised WSD methods.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Text Analysis; Language Parsing and Understanding

General Terms: Artificial Intelligence

1. INTRODUCTION

The work presented in this paper is aimed at developing and extending machine readable dictionary (MRD) based word sense disambiguation (WSD) techniques. The goal of WSD is to link ambiguous occurrences of the words in specific contexts to their meanings, as defined in a predefined sense inventory. For instance, given the following input:

(1) おとなしい 犬 を 飼い たい
otonashī *inu* *o* *kai* *tai*
quiet dog ACC keep want to
"(I) want to keep a quiet dog"

Baldwin, Timothy, Su Nam Kim, Francis Bond, Sanae Fujita, David Martinez and Takaaki Tanaka (to appear) A Reexamination of MRD-based Word Sense Disambiguation, ACM Transactions on Asian Language Information Processing, 9(1).

we aim to identify each content word (namely the first, second and fourth words: *otonashī*, *inu* and *ka(u)*,[1] respectively) as occurring with the sense corresponding to the provided English glosses. In the case of the second word (犬 [*inu*]), e.g., the dictionary lists two senses for the word, corresponding to the English "dog" and "spy", respectively.[2] We need to disambiguate the word to the first sense ("dog") in the given input. Similarly, we would identify the range of senses for all other content words (namely the first and fourth words), and disambiguate each relative to the context provided.

The relevance of MRDs to the task of WSD is that we build our method solely off information available in this pre-existing resource, namely using the textual definitions of each sense, and potentially ontological links in the dictionary. MRD-based WSD methods have the advantage that they are easily adaptable to any given MRD, and don't require any external data or additional sense annotations. That is, they provide the means to perform rapid development of WSD capabilities for languages or domains where sense-annotated data is not readily available.

There is a variety of reasons why we may wish to build a WSD system. For example, WSD has been shown to enhance parsing accuracy [Fujita et al. 2007; Agirre et al. 2008], so it could be for purely utilitarian reasons. Alternatively, we may require explicit word sense predictions, e.g. in a language understanding task [Bond et al. 2004] or crosslingual word glossing task [Yap and Baldwin 2007], presenting context-sensitive glosses of words in running text. For language understanding, there would be particular emphasis on predicting the most-likely sense for each word with high accuracy, whereas with word glossing, we could instead focus on filtering out *im*plausible senses and possibly retain multiple senses in the case that there is no standout single most-likely sense.

The construction of WSD systems has been the goal of many research initiatives (see for instance Ide and Véronis [1998] and Agirre and Edmonds [2006]). A broad interest in WSD and derivative tasks, and the need for standardised evaluation platforms, led to the establishment of the Senseval and SemEval initiatives.[3] Two key evaluation tasks (across multiple languages) in each of the four iterations of Senseval/SemEval to date have been an ALL WORDS and LEXICAL SAMPLE task. The ALL WORDS task requires that all (content) words in a sample text be disambiguated, and the LEXICAL SAMPLE task requires that, for a small set of words and selection of sentences containing that word for each, only the word of interest be disambiguated, all relative to a predefined sense inventory.

WSD systems are generally classified according to the type of tasks they are able to perform. The most relevant factors are the coverage of the system (e.g. ALL WORDS vs. LEXICAL SAMPLE, à la the Senseval tasks), and the granularity of the senses they are able to distinguish (e.g. homographs vs. fine-grained senses). The appropriate coverage and granularity of a system is determined relative to the final application of the WSD and the availability of data. For instance, for most

[1] The base form of the verb *kai*.

[2] An alternative view of this second sense would be to consider it a metaphoric extension of the basic "dog" sense, and lump the two senses together. We present it as a distinct sense here as that is how it is categorised in Lexeed, and hence in Hinoki.

[3] `http://www.senseval.org`

applications, such as the cross-lingual glossing example above, we would require full coverage of the text, but the granularity could be reasonably coarse.

WSD systems can be further classified according to the knowledge sources they use to build their models. A top-level distinction is made between supervised and unsupervised systems. The former rely on training instances that have been hand-tagged in order to build their classification models, while the latter build their classifiers relying on other types of knowledge, such as lexical databases or untagged corpora. The Senseval/SemEval evaluation tracks have shown that supervised systems perform better when training data is available, but they do not scale well to all words in context. Estimations of the effort required to build training data of reasonable size for all words (and all languages) are pessimistic [Mihalcea and Chklovski 2003]. This is known as the knowledge acquisition bottleneck, and is the main motivation behind research on unsupervised techniques.

While sense-annotated data is hard to come by, in recent years rich lexical resources have become increasingly available for a variety of languages. These knowledge sources are designed to be applied in language technology (LT) applications, and are often reusable for different tasks. In this paper, we aim to exploit an existing lexical resource to build an all-words Japanese word-sense disambiguator, that in turn can be applied to applications such as word glossing, or alternatively interfaced with downstream processing such as parse selection. The resource in question is the Hinoki Sensebank [Tanaka et al. 2006; Bond et al. 2006] and consists of the 28,000 most familiar words of Japanese, each of which is annotated with one or more basic senses from the Lexeed dictionary [Kasahara et al. 2004]. The senses take the form of a dictionary definition composed from only the 28,000 words contained in the dictionary, each of which is further manually sense annotated according to the Lexeed sense inventory. Hinoki also has a semi-automatically constructed ontology [Nichols et al. 2005]. This combination of features makes it an ideal resource for developing and evaluating WSD systems.

Through the Hinoki Sensebank, we investigate a number of areas of general interest to the WSD community. First, we test extensions to the Lesk algorithm [Lesk 1986; Banerjee and Pedersen 2003], focusing specifically on the impact of the scoring metric, the segment representation, and expansion via ontological semantics on WSD performance. These results are significant in porting the extended Lesk method to a language other than English, with the intention of improving the accuracy of unsupervised MRD-based WSD; this is particularly relevant for languages that do not have sense-disambiguated corpora. Second, we propose further extensions of the Lesk algorithm that make use of disambiguated definitions. In this, we shed light on the relative benefits we can expect from hand-tagging dictionary definitions, i.e. in introducing "semi-supervision" to the disambiguation task (c.f. Niu et al. [2005] and Pham et al. [2005]). The results are equally of interest for English, as the Extended WordNet [Harabagiu et al. 1999] provides an analogous resource with (semi-automatically) sense-annotated and parsed definitions which could be combined with this algorithm, and the Princeton WordNet Gloss Corpus includes fully sense-annotated definitions.[4]

[4] http://wordnet.princeton.edu/glosstag.shtml

In the remainder of this paper, we first summarise related work in unsupervised WSD in Section 2. Then we describe the basic method, resources, and characteristics of the Japanese language required to understand our work in Section 3. The extensions to the Lesk method are described in Section 4. Next, we present the evaluation of the different techniques in Section 5, and finally we discuss our conclusions in Sections 6 and 7.

2. RELATED WORK

As pointed out above, our work focuses on unsupervised and semi-supervised methods that target all words and parts of speech (POS) in context. We blur the supervised/unsupervised boundary somewhat in combining the basic unsupervised methods with hand-tagged definitions from Hinoki, in order to measure the improvement we can expect from sense-tagged dictionary data. Having said this, it is important to recognise that our method differs from conventional supervised methods in that it doesn't hinge on the availability of annotated training examples for every sense in the sense inventory. Hand-tagged definitions are less costly to produce than sense-annotated open text because: (1) the effects of discourse are limited, (2) the syntax is relatively simple, (3) there is significant semantic priming relative to the word being defined, and (4) there is generally explicit meta-tagging of the domain in technical definitions. In our experiments, we will make clear when sense-annotated data is being used.

Unsupervised methods often draw on a range of knowledge sources to build their models. Primarily the following types of lexical resources have been used for WSD: MRDs, lexical ontologies, and untagged corpora (monolingual corpora, second language corpora, and parallel corpora). Although early approaches focused on exploiting a single resource [Lesk 1986], recent trends show the benefits of combining different knowledge sources, such as multiple preprocessors [Stevenson 2003], or hierarchical relations from an ontology and untagged corpora [McCarthy et al. 2007]. In this summary, we will focus on a few representative systems that make use of different resources, noting that this is an area of very active research which we cannot do true justice to within the confines of this paper.

The Lesk method [Lesk 1986] is an MRD-based system that relies on counting the overlap between the words in the target context and the dictionary definitions of the senses.[5] In spite of its simplicity, it has been shown to be a hard baseline for unsupervised methods in Senseval [Kilgarriff and Rosenzweig 2000; Mihalcea et al. 2004; Pradhan et al. 2007]. Banerjee and Pedersen [2003] extended the Lesk method for WordNet-based WSD tasks, to include hierarchical data from the WordNet ontology [Fellbaum 1998] via an implicit representation of word context. They observed that the hierarchical relations significantly enhance the basic model. Both these methods will be described extensively in Section 3, as our approach is based on them.

A recent approach that combines ontological relations and untagged corpora was presented by McCarthy et al. [2007], who implemented an algorithm to automatically rank word senses in relation to a corpus. They relied on a thesaurus automat-

[5]See Mihalcea [2006] for a general overview of the Lesk method and various extensions people have made to it.

ically created using the method of Lin [1998], and then overlaid WordNet similarity scores onto this thesaurus in order to create the ranking. The last step was to assign to each test instance the sense of the highest-scoring sense in the ranking.

There are other techniques that exploit untagged data in order to build sense-tagged resources automatically. One such approach was first presented by Leacock et al. [1998], and follows these basic steps: (i) select a set of monosemous words that are related to the different senses of the target word, (ii) query the web to obtain examples for each relative, (iii) create a collection of training examples for each sense, and (iv) use a machine learning algorithm trained on the acquired collections to tag the test instances. This method has been used to bootstrap large sense-tagged corpora [Mihalcea 2002; Agirre and Martinez 2004].

Bootstrapping techniques consist of algorithms that learn from a few instances of labeled data (seeds) and a big set of unlabelled examples. In his widely-cited work, Yarowsky [1995] applied an iterative bootstrapping process to induce a classifier based on decision lists. With a minimal set of seed examples, disambiguation results comparable to supervised methods were obtained, although this was over a limited set of coarse-grained binary sense distinctions; this work has not been extended to fine-grained senses.

Parallel corpora have also been used to avoid the need for hand-tagged data. Following this approach, Chan and Ng [2005] built a classifier from English–Chinese parallel corpora. They considered English words which correspond to the same Chinese translation to have the same sense, and thus obtained sense-disambiguated data without manual annotation. Others who have used parallel corpora for WSD include Dagan and Itai [1994] and Diab and Resnik [2002]. However, large-scale parallel corpora are expensive to build, tend to have limited coverage, and only exist for a very small number of language pairs.

3. BACKGROUND

As background to our work, we first describe the lexical resources we used in our experiments, then outline aspects of Japanese relevant to this work, and finally present the basic and extended Lesk algorithms that are at the core of our approach.

3.1 The Hinoki Sensebank

All our experimentation is based on the Hinoki Sensebank [Tanaka et al. 2006]. The Hinoki Sensebank consists of sense annotations based on the Hinoki dictionary, which contains 28,000 high-familiarity Japanese words with 46,000 different senses [Kasahara et al. 2004]. The Hinoki dictionary is based on Kindaichi and Ikeda [1988], a standard monolingual dictionary of Japanese which has been repurposed for WSD.

The sense granularity of Hinoki is relatively coarse for most words, with the possible exception of light verbs, making it well suited to open-domain applications. The definition sentences for each sense were rewritten to use only the closed vocabulary of 28,000 words in the Hinoki lexicon (in addition to some function

[6]犬 $_1$ *inu* is also the hypernym for the following words: 愛犬 $_1$ [*aikeN*] "pet dog", 飼犬 $_1$ [*kaiinu*] "pet dog", カメ $_2$ [*kame*] "European dog", 狂犬 $_1$ [*kyōken*] "mad dog", 番犬 $_1$ [*baNkeN*] "guard dog", ...

INDEX	犬 *inu*	
POS	noun	LEXICAL-TYPE `noun-lex`
FAMILIARITY	6.53 [1–7]	FREQUENCY 67 ENTROPY 0.03
SENSE 1 0.99	DEFINITION	犬 1 科 の 食肉 1 動物 1 。 A carnivorous animal of the canidae family .
		家畜 1 として 古く 1 から 飼わ 1 れ 、 飼い主 1 に 忠実 1 。 Kept domestically from ancient times; loyal to their owners.
	EXAMPLE	犬 1 を 飼っ 1 て いる 家 3 が 多い 1 。 There are many households that keep dogs.
	HYPERNYM	動物 1 *dōbutsu* "animal"
	SEM. CLASS	⟨`537:beast`⟩ (⊂ ⟨`535:animal`⟩)
	WORDNET	*dog*1
SENSE 2 0.01	DEFINITION	警察 1 など の 回し者 1 。 スパイ 1 。 A secret agent for the police, etc. A spy.
	EXAMPLE	警察 1 の 犬 2 だけ に は 成り 4 たく ない 。 I want to turn into anything but a police spy.
	HYPERNYM	回し者 1 *mawashimono* "secret agent"
	SYNONYM	スパイ 1 *supai* "spy"
	SEM. CLASS	⟨`317:spy`⟩ (⊂ ⟨`317:spy`⟩)
	WORDNET	*spy*1

Fig. 1. First two senses for the word 犬 [*inu*] "dog" (with English glosses)[6]

POS	NOUN	VERB	ADJ	ADV
# Word Senses	2.86	3.65	3.58	3.08
% Monosemous	62.9	34.0	48.3	46.4
Agreement (Token)	0.803	0.772	0.770	0.648
Agreement (Type)	0.851	0.844	0.810	0.833

Table I. Average polysemy, relative occurrence of monosemous words, and sense-tagging agreement in Hinoki

words). Additionally, an example sentence was manually constructed to exemplify each of the 46,000 senses, once again using the closed vocabulary of the Hinoki dictionary. Both the definition sentences and example sentences were then manually sense-annotated by 5 native speakers of Japanese, from which a majority sense was extracted. There were 199,268 ambiguous tokens in the definition sentences. The average annotation rate was 1,799 tokens/day; disambiguating the definition sentences took 111 person days/annotator. Table I details the average polysemy and relative occurrence of monosemous words of different POS type in Hinoki, as well as inter-annotator agreement statistics [Bond et al. 2006].

We give a (slightly shortened) example of the entry for 犬 [*inu*] "dog" in Figure 1. The sense ID of each content word is indicated by a subscript, where the sense IDs have been ranked by their frequency in the Hinoki (and ID 1 thus indicates the

first sense for that word). In the case of 犬 [*inu*], there are two senses; of the 67 occurrences of 犬 [*inu*] in the Hinoki Sensebank, 99% (66/67) correspond to sense ID 1, and only 1% (1/67) to sense ID 2. The semantic class links to Goitaikei [Ikehara et al. 1997], and the WordNet ID to version 2.0 of WordNet [Fellbaum 1998]; neither of these resources are used in this research.

Subsequent to the development of the Hinoki Sensebank, an ontology was semi-automatically induced by parsing the first definition sentence for each sense [Bond et al. 2004; Nichols et al. 2005]. Hypernyms were determined by identifying the highest scoping content predicate (i.e. the genus). Other relation types such as synonymy and domain were also induced based on trigger patterns in the definition sentences. Because each word is sense tagged, the induced ontological relations link senses rather than just words.

For example, for 犬 $_1$ "inu", the hypernym is 動物 $_1$ [*dōbutsu*] "animal", as this was the highest scoping real predicate in the first sentence of the definition: "A carnivorous animal of the canidae family". 犬 $_1$ "inu" itself is the hypernym for many other word senses. We treat hypernymy and hyponymy as transitive equivalents, resulting in those senses which are identified as having 犬 $_1$ "inu" as their hypernym (e.g. 愛犬 $_1$ [*aikeN*] "pet dog" and 番犬 $_1$ [*baNkeN*] "guard dog"), are equivalently hyponyms of 犬 $_1$.

In 犬 $_2$ "inu", スパイ $_1$ [*supai*] "spy" is the only content predicate in the definition, and we therefore take it to be a synonym. This entry has no direct hyponyms.

In the version of the Hinoki Sensebank used in this research, there are 50,562 hypernym links, 765 hyponym links, 1,854 domain links, 682 meronym links and 14,287 synonym links (including abbreviations and nicknames).

The Hinoki Sensebank also contains annotated newspaper text, as well as syntactic and structural semantic annotations [Tanaka et al. 2006; Bond et al. 2006], although we do not use them in these experiments.

3.2 Peculiarities of Japanese

While the basic method we propose is language independent, all our experiments are targeted exclusively at Japanese. As such, there are a number of features of Japanese which impinge on our experiments, as outlined below.

First, Japanese is made up of 3 basic alphabets: hiragana, katakana (both syllabic in nature) and kanji (logographic in nature).[7]

Second, Japanese is a non-segmenting language, i.e. there is no explicit orthographic representation of word boundaries. The native rendering of (1), e.g., is おとなしい犬を飼いたい. Fortunately, various packages exist to automatically segment Japanese strings into words and also POS tag and lemmatise the words. All the Hinoki data has been pre-segmented using ChaSen [Matsumoto et al. 2003], and we make use of this in our experiments. It is also possible, however, to sidestep segmentation altogether and represent the string via its component characters, under the expectation that the predominance of logographic kanji characters will provide a form of semantic "smoothing".

Third, Japanese has relatively free word order, or strictly speaking, word order within phrases is largely fixed but the relative ordering of phrases governed by

[7]Modern Japanese also standardly incorporates Arabic numerals and the Latin alphabet.

Algorithm 1 Generalised Lesk algorithm

for each word $w_i \in$ context $\mathbf{w} = w_1 w_2 ... w_n$ **do**
  for each sense $s_{i,j}$ and definition $\mathbf{d}_{i,j}$ **do**
    $score(s_{i,j}) = similarity(\mathbf{w}, \mathbf{d}_{i,j})$
  end for
  $s_i^* = \arg\max_j score(s_{i,j})$
end for

a given predicate is relatively free. For the purposes of WSD, this motivates a representation of context which captures local word order but abstracts away from phrase order.

3.3 Basic Lesk

The original Lesk algorithm [Lesk 1986] performs WSD by calculating the relative word overlap between the context of usage of a target word, and the dictionary definition[8] of each of its senses in a given MRD. The sense with the highest overlap is then selected as the most plausible hypothesis. This general method can be formalised according to Algorithm 1, which also forms the basis of our proposed method.

To see the basic Lesk algorithm in action, let us take (1) as input. Assuming a POS tagging and lemmatisation step,[9] the context of 犬 "dog" in (1) would be represented as the following bag of lemmatised content words (see Section 4.2 for details):

(1') { おとなしい , 飼う }

Further assuming the following (multi-sentence) definition for the first sense of 犬 (i.e. "dog"):

(2) 犬 科 の 食肉 動物 。 家畜 として
inu *ka* *no* *shokuniku* *dōbutsu* . *kachiku* *toshite*
canidae family GEN carnivore animal . domestic as
古く から 飼わ れ 、 飼い主 に 忠実 。
furuku *kara* *kawa* *re* , *kainushi* *ni* *chūjitsu* .
ancient times from keep PASS , owner DAT loyal .
"A carnivorous animal of the canidae family. Kept domestically from ancient times; loyal to their owners."

we would generate the following bag of (content) lemmas:

(2') { 犬 , 食肉 , 動物 , 家畜 , 古い , 飼う , 飼い主 , 忠実 }

Similarly, given the following definition for the second sense of 犬(i.e. "spy"):

[8] In Lesk [1986] and Banerjee and Pedersen [2003], these definitions are referred to as "glosses". In this paper, we reserve the word "gloss" to refer to the word translations for a target word.
[9] Both based on ChaSen [Matsumoto et al. 2003].

(3) 警察 など の 回し者 。 スパイ 。
keisatsu *nado* *no* *mawashimono* . *supai* .
police etcetera GEN secret agent . spy .
"A secret agent for the police, etc. A spy."

we would generate the following bag of (content) lemmas:

(3') { 警察 , 回し者 , スパイ }

In the original Lesk algorithm, the *similarity* calculation in Algorithm 1 takes the form of a simple set intersection operation and determination of the cardinality of the resulting set. In our example, the algorithm would correctly predict that (1) contains an instance of the first sense of *inu* due to the definition containing one lemma in common with the context (namely 飼う "keep"), as compared to the second sense where there are no overlapping lemmas.

3.4 Extended Lesk

An obvious weakness of the original Lesk algorithm is that it requires that the exact words used in the definitions be included in each usage of the target word. To redress this shortcoming, Banerjee and Pedersen [2003] extended the basic algorithm for WordNet-based WSD tasks to include ontological semantics, i.e. expanding the method to compare not only the definition of a given word sense, but also the definitions of word senses ontologically related to it (e.g. through hypernymy, hyponymy or meronymy).[10] Further, they modified the original algorithm to compare the definitions of each sense of each word within n words of the target word, rather than comparing the words in the context of the target word directly with the definition of each target word sense. To return to our example from above, given (1) and the target word *inu*, the Banerjee and Pedersen [2003] algorithm would identify *otonashī* and *kau* as context words of the target word, and identify all senses of each of the context words in Hinoki. It would further identify directly related senses of each sense of both *inu* and the context words (i.e. senses which are directly linked by a single edge). Finally, for each each sense of *inu*, it would take first the expanded set of that sense and all ontologically-linked senses, and second the expanded set of all senses of context words and their ontologically-linked senses, and perform a comparison of the definitions in the pairwise cross-product of the two sets, returning the combined sum of all such comparisons.

Formally, Banerjee and Pedersen [2003] formulate their *similarity* calculation as follows for a given word sense pairing (A, B):

$$similarity(A, B) = \sum_{\langle R_i, R_j \rangle \in \text{RELPAIRS}} score(R_i(A), R_j(B)) \quad (4)$$

where *score* is calculated as the square of the word length of the longest common sub-string in the definitions of $R_i(A)$ and $R_j(B)$, and RELPAIRS is a set of relation pairs, e.g., $\{\langle def, def \rangle, \langle hype, def \rangle, \langle def, hype \rangle\}$. Note here that the method

[10] In earlier work [Baldwin et al. 2008], we experimented with a predecessor method, also proposed by Banerjee and Pedersen [2002], and hence present slightly different results to those presented in this paper.

doesn't perform explicit comparison of the target word context with the (extended) word definition, but instead compares the (extended) word definition with the (extended) definitions of context words. In this research, we define RELPAIRS to be $\{\langle def, def\rangle, \langle hype, hype\rangle, \langle hypo, hypo\rangle\}$, indicating that we compare the definitions of the two senses, the definitions of the hypernyms of the two senses (each of which is a set in itself), and the definitions of the hyponyms of the two senses (each of which is, again, a set in itself).

4. PROPOSED ALGORITHM

The basis of our algorithm is an amalgam of the original Lesk algorithm and the extended algorithm of Banerjee and Pedersen [2003]: (1) like the original algorithm, we compare the context of the target word directly with definitions, unlike Banerjee and Pedersen who only ever compare definition sentences; and (2) like Banerjee and Pedersen we extend the original algorithm by looking beyond the definitions of the individual senses to include the definitions of ontologically-related senses, but our approach for doing this is to take the union of multisets associated with each definition rather than combine pairwise comparisons of individual definitions.

Below, we outline the specifics of the proposed algorithm, in terms of the scoring mechanism, tokenisation strategy, and different strategies for expanding the definitions.

4.1 Scoring Mechanism

In Algorithm 1, *similarity* provides the means to score a given pairing of context $\mathbf{w}$ and definition $\mathbf{d}_{i,j}$. In the original Lesk algorithm, *similarity* was simply the number of words in common between the two, which Banerjee and Pedersen [2003] modified by squaring the size of the longest overlapping sub-string. While squaring is well motivated in terms of preferring longer substring matches, longer definitions are preferred as they are a priori more likely to generate longer matches; the calculation of the longest substring is also computational expensive. We thus adopt a cheaper scoring mechanism which is normalised relative to the length of $\mathbf{w}$ and $\mathbf{d}_{i,j}$, but ignores the original order of the segments. Namely, we use the Dice coefficient, defined for sets A and B as follows:

$$sim_{\text{DICE}}(A, B) = \frac{2\,|A \cap B|}{|A| + |B|} \tag{5}$$

4.2 Tokenisation

As mentioned in Section 3.2, Japanese is a non-segmenting language which makes heavy use of logographic characters. As such, as an alternative to segmentation via a word splitter such as ChaSen and using words (or lemmas) as our segment unit, we can simply tokenise strings into their component characters and use characters as our segments. Note that independent of the choice of segment granularity, we require word splitting in order to identify the words in a string to look up senses for in our dictionary; that is, we are not doing away with the need for word segmentation in adopting a character indexing strategy, we are simply modifying the underlying string representation which our *similarity* mechanism is applied to. Character and word tokenisation have been compared in the context of Japanese information

retrieval [Fujii and Croft 1993] and translation retrieval [Baldwin 2001], and in both of these tasks, characters have been found to be the superior representation in certain contexts. It is thus interesting to investigate whether these findings carry across to WSD.

4.3 Expanded Definitions

The main direction in which Banerjee and Pedersen [2003] successfully extended the Lesk algorithm was in including ontologically-linked definitions (e.g. hyponyms and hypernyms). In the method of Banerjee and Pedersen, the expansion takes the form of carrying out multiple comparisons of different pairings of definitions and summing the individual scores together. Our approach is instead to use ontological links to first expand the definition for each word sense, and carry out a single comparison of the expanded definition with the context string of the target word. In information retrieval terms, therefore, our method equates to query expansion.

We experiment with a number of approaches to expansion. First, we expand the definition sentences to in turn include the definitions of each word contained in that definition. In the case of the first sense of *inu*, for example, in (2), we expand the definition to include the definitions for each word in the definition which is also defined in the Hinoki Sensebank, namely 犬 [*inu*], 食肉 [*shokuniku*], 動物 [*dōbutsu*], and so on.[11] As all of the definitions are sense-tagged, this expansion can occur either in a sense-sensitive manner, expanding the definition to include only the definitions of the sense instantiated in the definition, or a sense-insensitive manner, expanding the definition to include all senses of each word.

Second (and orthogonally), we follow Banerjee and Pedersen [2003] in expanding the definitions to include words from the definitions for the synonyms, hypernyms and/or hyponyms of a given sense of the target word.[12] For example, when we combine the first sense of *inu* (Fig 1) with its hypernym *dōbutsu* (Fig 2), we add the set of content words in $dōbutsu_1$'s definition { 生物 , 大きな , 区分 } to those of inu_1 { 犬 , 食肉 , 動物 , 家畜 , 古い , 飼う , 飼い主 , 忠実 } to give the set: { 犬 , 食肉 , 動物 , 家畜 , 古い , 飼う , 飼い主 , 忠実 , 生物 , 大きな , 区分 }. We then use this set to match against the context of the word being disambiguated. As with the expansion of definitions via the definitions of their component words, this ontology-based expansion can take place in either a sense-sensitive manner (i.e. expanding out only the synonyms, hyponyms and hypernyms of a given sense) or sense-insensitive manner (i.e. expanding out all synonyms, hyponyms and hypernyms for all senses of a given word). Note that we use only one level of lexical relations, i.e. only direct hyponyms and hypernyms.

[11] Note that in this case the definition includes the target word (*inu*), meaning that we duplicate all segments in the definition.

[12] Note that Banerjee and Pedersen [2003] did not experiment with synonyms, as WordNet provides definitions at the synset level, i.e. for a cluster of synonyms, such that synonym-to-synonym definition comparison will always result in a full string match. They did, however, make use of the richer set of ontological relations provided in WordNet, including meronymy, holonymy, also-see, and attribute, all of which we ignore.

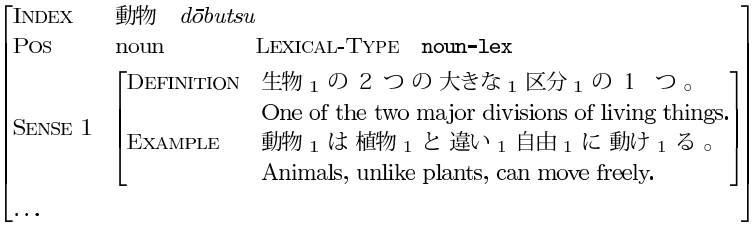

Fig. 2. Simplified first sense for the word 動物 [*dōbutsu*] "animal" (with English glosses)

4.4 Other Extensions

We also experimented with a number of other extensions to the proposed method, but found them to have relatively little impact so don't present detailed results here. These extensions included:

—POS-based filtering of the tokens for word-based tokenisation (e.g. treating verbal and demonstrative instances of the word ある [*aru*] as distinct tokens); this was found to have very little impact on results, due to very few POS-differentiated homographs in Japanese

—Stop word filtering, using the same set of stop words as was used in Nichols et al. [2005]; this led to very slight increments in accuracy across the board (of the order of 0.001)

—Different n-gram orders for both words and characters; this was found to improve results very marginally for character-based tokenisation in isolated cases, but generally drove results down

—Different scoring metrics (e.g. cosine similarity and simple set overlap), but the Dice coefficient was found to be the best overall.

5. EVALUATION

We evaluate our various extensions over two datasets: (1) the example sentences in the Hinoki Sensebank, and (2) the Senseval-2 Japanese dictionary task [Shirai 2002].

Each sense in the Hinoki Sensebank has one example sentence, and therefore we are guaranteed to have at least one token instance of every sense in the sensebank. As stated above, all content words in the example sentences were sense annotated [Tanaka et al. 2006].

The Senseval-2 Japanese dictionary task is a LEXICAL SAMPLE task, which used the Iwanami Kokugo Jiten as its original sense inventory. We include this dataset for calibration purposes, in comparing our method with previously published results on Japanese WSD. As we do not have access to an ontology or sense-annotated definitions for the Iwanami Kokugo Jiten, however, we are not able to run our

method over it directly. Instead, we used a re-annotated version of the Senseval-2 data based on Hinoki senses. Similarly to the example sentences, the Senseval-2 data was 5-way manually sense annotated, and sense-arbitrated according to the majority annotation. Naturally, the sense granularity of Hinoki is not identical to that for the Iwanami Kokugo Jiten, and hence the sense distribution of our re-annotated data diverges from that for the original data. Fortunately, all target words in the Senseval-2 task were common enough that Hinoki had full coverage over them.

All results below are reported in terms of simple accuracy.

For the two datasets, we use two baselines: a random baseline and the first-sense baseline. The former assigns one of the candidate senses randomly, and the latter always picks sense ID 1 for each word (i.e. the word sense with the most token occurrences within the Hinoki definitions). As such, random sense assignment is an unsupervised baseline, while first-sense assignment is a supervised baseline. We will use these values as a reference for our algorithm. Note that the (supervised) first-sense baseline has been shown to be hard to beat for unsupervised systems [Kilgarriff 2004; McCarthy et al. 2007], for example for 犬 [*inu*], the first sense baseline has an accuracy of 99% over the 67 occurrences of 犬 in the sensebank (both definition and example sentences).

As a benchmark MRD-based WSD method, we reimplemented the original Banerjee and Pedersen [2003] method to work over the Hinoki data. As mentioned above, our reimplementation is slightly different to that applied to the English WordNet data, in terms of the selection of ontological relations we use.

5.1 Hinoki Example Sentences

The goal of these experiments is to tag all the words that occur in the example sentences in the Hinoki Sensebank.

In our first sub-experiment, we experiment with the two tokenisation strategies (characters [CHAR] vs. words [WORD]) and the optional use of extended definitions (without ontological relations, and at the word rather than sense level). The results are presented in the top portion of Table II.

The first finding is that characters are in all cases superior to words as our segment granularity. This is a somewhat surprising result, given that the median word length is 2 characters, based on which it would not appear that words would generate excessive data sparseness.

Extended definitions are also shown to be superior to simple definitions, although the relative increment in making use of large amounts of sense annotations is lesser than that of characters vs. words.

Note that at this point, our best-performing method is at the level of the unsupervised (random) baseline, and well below the supervised (first sense) baseline.

Having found that extended definitions improve results to a small degree, we next experiment with the inclusion of various ontological relations to expand the original definitions. Here, we persevere with the use of word and characters, and experiment with the addition of synonyms, hypernyms and/or hyponyms, with and without the extended definitions (we don't present results or all combinations of lexical relations for reasons of space). We further test the impact of the sense annotations, in rerunning our experiments with the ontology in a sense-*insensitive*

	Sense-sensitive		Sense-insensitive	
	Word	Char	Word	Char
unsupervised (random) baseline:			0.527	
supervised (first-sense) baseline:			0.633	
Banerjee and Pedersen [2003]			0.563	
simple	—	—	0.469	0.524
+extdef	—	—	0.489	0.527
+syn	0.560	0.538	0.548	0.543
+hyper	0.559	0.539	0.548	0.537
+hypo	0.656	0.644	**0.655**	**0.644**
+hyper +hypo	0.648	0.641	0.629	0.630
+syn +hyper +hypo	0.650	0.633	0.627	0.623
+extdef +syn	0.577	0.560	0.551	0.543
+extdef +hyper	0.577	0.563	0.551	0.542
+extdef +hypo	0.653	0.646	0.649	**0.644**
+extdef +hyper +hypo	**0.683**	**0.671**	0.631	0.627
+extdef +syn +hyper +hypo	0.680	0.661	0.632	0.621

Table II. Accuracy using ontology-based definition extension (with/without word sense information) and words (Word) or characters (Char) as our segment unit (best-performing method in each column in **boldface**)

manner, i.e. by adding in the union of word-level hypernyms and/or hyponyms. The results are described in the bottom portion of Table II.

Adding in the ontology makes a considerable difference to our results, in line with the findings of Banerjee and Pedersen [2003]. Hyponyms are better discriminators than hypernyms (assuming a given word sense has a hyponym — the Hinoki ontology is relatively flat), partly because while a given word sense will have (at most) one hypernym, it often has multiple hyponyms (if any at all). Adding in hypernyms or hyponyms, in fact, has a greater impact on results than simple extended definitions (+extdef), especially for the word-based representation. The best overall results are produced for the combination of all ontological relations (i.e. extended definitions, hypernyms and hyponyms), achieving an accuracy level above both the unsupervised (random) and supervised (first-sense) baselines and the Banerjee and Pedersen [2003] method.

Looking at the sense-insensitive results, simple hyponyms (without extended definitions) and word-based tokenisation produced the best results out of all the variants tried, at an accuracy of 0.655, once again above both baselines and the original Banerjee and Pedersen [2003] method. This compares with an accuracy of 0.683 achieved for the best of the sense-sensitive methods, indicating that sense information in the definitions enhances WSD performance. This reinforces our expectation that richly annotated lexical resources improve performance, but also indicates that in the absence of sense annotations in the lexical resource, our method is still able to achieve highly competitive results. With richer information to work with, character-based methods uniformly give worse results.

In Table III, we present a breakdown of the results over the four main word

		ALL	NOUN	VERB	ADJ	ADV
UNSUPERVISED (RANDOM) BASELINE:		0.527	0.641	0.252	0.415	0.564
SUPERVISED (FIRST-SENSE) BASELINE:		0.633	0.718	0.407	0.547	0.628
Banerjee and Pedersen [2003]		0.563	0.672	0.290	0.440	0.540
WORD	simple	0.469	0.620	0.145	0.294	0.388
	+extdef	0.489	0.630	0.179	0.306	0.451
	+syn	0.560	0.679	0.281	0.420	0.609
	+hyper	0.559	0.679	0.281	0.384	0.609
	+hypo	0.656	0.747	0.432	0.571	0.645
	+hyper +hypo	0.648	0.739	0.423	0.553	0.653
	+syn +hyper +hypo	0.650	0.743	0.419	0.615	**0.665**
	+extdef +syn	0.577	0.717	0.282	0.315	0.590
	+extdef +hyper	0.577	0.717	0.282	0.380	0.590
	+extdef +hypo	0.653	0.741	**0.434**	0.584	0.664
	+extdef +hyper +hypo	**0.683**	**0.789**	0.429	0.574	0.644
	+extdef +syn +hyper +hypo	0.680	0.785	0.428	**0.619**	0.659
CHAR	simple	0.524	0.652	0.235	0.365	0.515
	+extdef	0.527	0.641	0.263	0.402	0.575
	+syn	0.538	0.666	0.263	0.301	0.575
	+hyper	0.538	0.666	0.267	0.382	0.572
	+hypo	0.644	0.729	0.431	0.573	0.639
	+hyper +hypo	0.640	0.728	0.423	0.564	0.636
	+syn +hyper +hypo	0.633	0.722	0.413	0.590	0.635
	+extdef +syn	0.559	0.705	0.259	0.261	0.565
	+extdef +hyper	0.563	0.709	0.262	0.360	0.567
	+extdef +hypo	0.646	0.737	0.421	0.574	0.642
	+extdef +hyper +hypo	0.671	0.776	0.421	0.564	0.632
	+extdef +syn +hyper +hypo	0.660	0.763	0.414	0.594	0.649

Table III. Breakdown of results across word classes using ontology-based definition extension (with word sense information) and word (WORD) and character (CHAR) unigrams

classes (nouns, verbs, adjectives and adverbs), with sense-sensitive lexical relation expansion. We observe a relatively consistent trend across the four word classes, in terms of words generally outperforming characters when we introduce the lexical relations, and the addition of extra lexical relations generally improving accuracy. We also observe a marked imbalance in both the baseline accuracies and performance levels across the different word classes. As commonly observed for ALL WORDS WSD tasks, verbs are the hardest POS to disambiguate and nouns the easiest. This is in terms of both the baselines and the relative boost in accuracy with our method (based on error rate reduction[13]). In all cases, the lexical relation which produces the singular greatest increment in accuracy is hyponyms, although the effect is considerably less pronounced for adverbs than the other three word classes. Overall, more lexical relations generally lead to higher accuracy, but there are subtle differences across the word classes in terms of exactly what combination of lexical relations produces the best accuracy overall.

[13]For a method with accuracy a relative to a baseline with accuracy b, the error rate reduction is

	Sense-sensitive		Sense-insensitive	
	Word	Char	Word	Char
Unsupervised (random) baseline:		0.310		
Supervised (first-sense) baseline:		0.577		
Banerjee and Pedersen [2003]		0.700		
simple	—	—	0.373	0.404
+extdef	—	—	0.362	0.420
+syn	0.696	0.689	0.691	0.685
+hyper	0.450	0.441	0.426	0.425
+hypo	0.577	0.568	0.610	0.616
+hyper +hypo	0.591	0.585	0.608	0.596
+syn +hyper +hypo	0.768	0.763	0.760	0.760
+extdef +syn	0.694	0.695	0.691	0.692
+extdef +hyper	0.484	0.451	0.432	0.371
+extdef +hypo	0.630	0.616	0.622	0.610
+extdef +hyper +hypo	0.624	0.624	0.602	0.593
+extdef +syn +hyper +hypo	**0.776**	**0.764**	**0.765**	**0.761**

Table IV. Accuracy over the Senseval-2 data

5.2 Senseval-2 Japanese Dictionary Task

In our second set of experiments we apply our proposed method to the Senseval-2 Japanese dictionary task [Shirai 2002] in order to calibrate our results against previously-published results for Japanese WSD. Recall that this is a lexical sample task, and that our evaluation is relative to Hinoki re-annotations of the same dataset, although the relative polysemy for the original data and the re-annotated version is largely the same [Tanaka et al. 2006]. The first sense baselines for the two sets of annotations differ significantly, however, with an accuracy of 0.726 reported for the original task, and 0.577 for the re-annotated Hinoki variant.

In Table IV, we present the results over the Senseval-2 data for the best-performing systems from our earlier experiments. As before, we include results over both words and characters, and with sense-sensitive and sense-insensitive ontology expansion.

Our results largely mirror those of Table II. The best methods surpass the random and first sense baselines, as well as Banerjee and Pedersen [2003]. In this case, the synonyms have the greatest impact out of the three lexical relations, but the combination of all three lexical relations is the best overall performer. There is less difference between word and character tokenisation than in our first experiment, and the relative impact of sense annotations was if anything even less pronounced than for the example sentence task.

We achieved our best accuracy with word tokenisation, extended definitions, all lexical relations and sense-sensitivity, at an accuracy of 0.776. This represents an error reduction rate of 47.0% over the first-sense baseline, and compares favourably with an error rate reduction of 21.9% for the best of the WSD systems in the original

calculated according to: $\frac{a-b}{1-b}$.

Senseval-2 task, based on "mixed-grain" senses [Kurohashi and Shirai 2001]. It is particularly impressive given that our method is semi-supervised while the Senseval-2 system is a conventional supervised word sense disambiguator. Even when we take away the sense-sensitivity, the error reduction rate drops back only fractionally to 44.4%.

In more recent work, Tanaka et al. [2007] use the Hinoki Sensebank to train a supervised classifier, which uses both the ontological information and semantic dependencies from a parser. The split into training and test is different from ours, so the results are not directly comparable. The most comparable results are where Tanaka et al. [2007] train on both definition and example sentences, and test on a held-out set of example sentences. Here, the first sense baseline is 0.704 (compared to 0.633 in our case, training on definition sentences and testing on example sentences), and the supervised system achieved an accuracy of 0.788, easily surpassing our best unsupervised results of 0.683 in terms of both raw accuracy and error rate reduction (28.4% vs. 13.6%).

6. DISCUSSION

Over the Hinoki example sentence data, the use of the sense-tagged data to link senses in the ontology gave the best result overall at an accuracy of 0.683. This was using extended definitions, hypernym and hyponym links, and word tokenisation. Without the sense information, the best result was 0.655 using only hyponym links. As can be expected, more precise information leads to a higher precision. The extra information does not, however, come for free: the ontology was automatically extracted, but underwent minor manual post-correction as errors were noticed, and the sense-links were manually annotated. If we were interested solely in WSD, the annotation could have been greatly reduced, as only the linked words (e.g. hypernyms, hyponyms and synonyms) need to be disambiguated, accounting for roughly one third of the total. Alternatively, the annotation could have been performed automatically, as was done for the first version of Extended WordNet [Harabagiu et al. 1999], or just the hypernyms disambiguated, as in Rigau et al. [1997]. In our opinion, the 110 person/days (per annotator) to sense tag the definition sentences was a small cost compared to the overall outlay to build the lexicon, and well justified. In the case of WordNet, all links are relative to synsets[14] and hence sense specified, so using WordNet links is equivalent to our sense-sensitive system.

Comparing our method to Banerjee and Pedersen [2003], not only did we achieve higher accuracy over two separate datasets, but our method is computationally cheaper, as we perform a single set intersection calculation rather than multiple calculations of the length of the longest common subsequence between glosses. It is important to point out that we did not attempt to optimise the parameterisation of the Banerjee and Pedersen [2003] method over either of our datasets, in terms of the combinations of lexical relations or size of the context window. In this sense, more research could be carried out to thoroughly explore the impact of the parameter setting on the accuracy of the method.

Of character and word tokenisation, word tokenisation was found to be slightly

[14]Strictly speaking, antonyms, e.g., are word- rather than sense-specific, but they are not used in this research or the work of Banerjee and Pedersen.

superior when combined with ontological expansion, but in the absence of an ontology, character tokenisation consistently outperformed word tokenisation. Combining this with the finding that the inclusion of ontological relations boosted accuracy considerably, the overall finding is that it is well worth inducing an ontology (e.g. using the automatic method of Nichols et al. [2005]), and that even in the absence of sense annotation, this will have the single greatest impact on the accuracy of WSD based on our proposed method.

7. CONCLUSION

We proposed a new method for MRD-based word sense disambiguation based on definition expansion via an ontology, building on the work of Lesk [1986] and Banerjee and Pedersen [2003]. In this, we experimented with character- and word-based tokenisation, definition extension based on the words in the original definition sentences, and a range of lexical relations. With the lexical relations, we experimented with both sense-sensitive and sense-insensitive expansion, interpreting the lexical relations as linking either word senses or words, respectively. In doing so, we measured the contribution of sense-tagged definitions to the overall disambiguation performance. We evaluated our proposed method over two Japanese datasets: example sentences from the Hinoki Sensebank, and a retagged version of the Senseval-2 Japanese dictionary task. In making maximal use of the available lexical relations and definition extension, we were able to surpass both unsupervised and supervised baselines for the two datasets, and the method of Banerjee and Pedersen [2003]. We also found that sensitising the lexical relations to word sense consistently improved the accuracy of our method, and that the method performed best over nouns.

One of the strengths of our method is that it can be applied equally to all words in running text, at an accuracy level higher than a first-sense baseline. This full-coverage system opens the way to further enhancements, such as the contribution of extra sense-tagged examples to the expansion, or the combination of different WSD algorithms. In future work, we intend to integrate the WSD tool with other Japanese text processing applications, such as a cross-lingual glossing tool for learners of Japanese text [Yap and Baldwin 2007] and parse selection for Japanese grammars [Fujita et al. 2007].

Acknowledgements

We wish to thank Hiromi Nakaiwa and Masaaki Nagata for their support and advice throughout this work. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

REFERENCES

Agirre, E., Baldwin, T., and Martinez, D. 2008. Improving parsing and PP attachment performance with sense information. In *Proceedings of the 46th Annual Meeting of the ACL: HLT*. Columbus, USA, 317–325.

Agirre, E. and Edmonds, P., Eds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer, Dordrecht, Netherlands.

Agirre, E. and Martinez, D. 2004. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*. Barcelona, Spain, 25–32.

Baldwin, T. 2001. Low-cost, high-performance translation retrieval: Dumber is better. In *Proceedings of the 39th Annual Meeting of the ACL and 10th Conference of the EACL (ACL-EACL 2001)*. Toulouse, France, 18–25.

Baldwin, T., Kim, S. N., Bond, F., Fujita, S., Martinez, D., and Tanaka, T. 2008. MRD-based word sense disambiguation: Further extending Lesk. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*. Hyderabad, India, 775–780.

Banerjee, S. and Pedersen, T. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*. Mexico City, Mexico, 136–145.

Banerjee, S. and Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003)*. Acapulco, Mexico, 805–810.

Bond, F., Fujita, S., Hashimoto, C., Kasahara, K., Nariyama, S., Nichols, E., Ohtani, A., Tanaka, T., and Amano, S. 2004. The Hinoki treebank: A treebank for text understanding. In *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*. Hainan Island, China, 554–559.

Bond, F., Fujita, S., and Tanaka, T. 2006. The Hinoki syntactic and semantic treebank of Japanese. *Language Resources and Evaluation 40,* 3–4, 253–261. (Special Issue on Asian Language Technology).

Bond, F., Nichols, E., Fujita, S., and Tanaka, T. 2004. Acquiring an ontology for a fundamental vocabulary. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*. Geneva, Switzerland, 1319–1325.

Chan, Y. S. and Ng, H. T. 2005. Scaling up word sense disambiguation via parallel texts. In *Proceedings of the 20th Annual Conference on Artificial Intelligence (AAAI-07)*. Pittsburgh, USA, 1037–1042.

Dagan, I. and Itai, A. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics 20,* 4, 563–596.

Diab, M. and Resnik, P. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the ACL and 3rd Annual Meeting of the NAACL (ACL-02)*. Pittsburgh, USA, 255–262.

Fellbaum, C., Ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

Fujii, H. and Croft, W. B. 1993. A comparison of indexing techniques for Japanese text retrieval. In *Proceedings of 16th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*. Pittsburgh, USA, 237–246.

Fujita, S., Bond, F., Oepen, S., and Tanaka, T. 2007. Exploiting semantic information for HPSG parse selection. In *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*. Prague, Czech Republic, 25–32.

Harabagiu, S. M., Miller, G. A., and Moldovan, D. I. 1999. WordNet 2 – a morphologically and semantically enhanced resource. In *Proceedings of SIGLEX99: Standardizing Lexical Resources*. College Park, USA, 1–8.

Ide, N. and Véronis, J. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics 24,* 1, 1–40.

Ikehara, S., Miyazaki, M., Yokoo, A., Shirai, S., Nakaiwa, H., Ogura, K., Ooyama, Y., and Hayashi, Y. 1997. *Nihongo Goi Taikei – A Japanese Lexicon*. Iwanami Shoten. 5 volumes. (in Japanese).

Kasahara, K., Sato, H., Bond, F., Tanaka, T., Fujita, S., Kanasugi, T., and Amano, S. 2004. Construction of a Japanese semantic lexicon: Lexeed. In *Proceedings of SIG NLC-159*. Tokyo, Japan, 75–82.

Kilgarriff, A. 2004. How dominant is the commonest sense of a word? Tech. Rep. ITRI-04-10, Information Technology Research Institute, University of Brighton.

KILGARRIFF, A. AND ROSENZWEIG, J. 2000. English Senseval: Report and results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*. Athens, Greece, 1239–1244.

KINDAICHI, H. AND IKEDA, Y. 1988. *Gakken Japanese Dictionary 2nd edition*. Gakken Co., Ltd.

KUROHASHI, S. AND SHIRAI, K. 2001. SENSEVAL-2 Japanese tasks. In *IEICE Technical Report NLC 2001-10*. 1–8. (in Japanese).

LEACOCK, C., CHODOROW, M., AND MILLER, G. A. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics 24,* 1, 147–165.

LESK, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*. Ontario, Canada, 24–26.

LIN, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*. Montreal, Canada, 768–774.

MATSUMOTO, Y., KITAUCHI, A., YAMASHITA, T., HIRANO, Y., MATSUDA, H., TAKAOKA, K., AND ASAHARA, M. 2003. *Japanese Morphological Analysis System ChaSen Version 2.3.3 Manual*. Tech. rep., NAIST.

MCCARTHY, D., KOELING, R., WEEDS, J., AND CARROLL, J. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics 4,* 33, 553–590.

MIHALCEA, R. 2002. Bootstrapping large sense tagged corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Spain, 1407–1411.

MIHALCEA, R. 2006. Knowledge-based methods. In *Word Sense Disambiguation: Algorithms and Applications*, E. Agirre and P. Edmonds, Eds. Springer, Dordrecht, Netherlands.

MIHALCEA, R. AND CHKLOVSKI, T. 2003. Open Mind word expert: Creating large annotated data collections with web users' help. In *Proceedings of the EACL 2003 Workshop on Linguistically Annotated Corpora (LINC 2003)*. Budapest, Hungary, 53–61.

MIHALCEA, R., CHKLOVSKI, T., AND KILGARRIFF, A. 2004. The Senseval-3 English lexical sample task. In *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain, 25–28.

NICHOLS, E., BOND, F., AND FLICKINGER, D. 2005. Robust ontology acquisition from machine-readable dictionaries. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-2005)*. Edinburgh, UK, 1111–1116.

NIU, Z.-Y., JI, D.-H., AND TAN, C. L. 2005. Word sense disambiguation using label propagation based semi-supervised learning. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, USA, 395–402.

PHAM, T. P., NG, H. T., AND LEE, W. S. 2005. Word sense disambiguation with semi-supervised learning. In *Proceedings of the 20th Annual Conference on Artificial Intelligence (AAAI-07)*. Pittsburgh, Pennsylvania, 1093–1098.

PRADHAN, S., LOPER, E., DLIGACH, D., AND PALMER, M. 2007. SemEval-2007 Task-17: English lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Prague, Czech Republic, 87–92.

RIGAU, G., ATSERIAS, J., AND AGIRRE, E. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the EACL (ACL-EACL'97)*. Madrid, Spain, 48–55.

SHIRAI, K. 2002. Construction of a word sense tagged corpus for SENSEVAL-2 Japanese dictionary task. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Spain, 605–608.

STEVENSON, M. 2003. *Word Sense Disambiguation: The Case for Combining Knowledge Sources*. CSLI Publications, Stanford, USA.

TANAKA, T., BOND, F., BALDWIN, T., FUJITA, S., AND HASHIMOTO, C. 2007. Word sense disambiguation incorporating lexical and structural semantic information. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2007 (EMNLP-CoNLL 2007)*. Prague, Czech Republic, 477–485.

TANAKA, T., BOND, F., AND FUJITA, S. 2006. The Hinoki sensebank — a large-scale word sense tagged corpus of Japanese —. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*. Sydney, Australia, 62–69.

YAP, W. AND BALDWIN, T. 2007. Dictionary alignment for context-sensitive word glossing. In *Proceedings of the Australasian Language Technology Workshop 2007*. Melbourne, Australia, 125–133.

YAROWSKY, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. Cambridge, USA, 189–196.