Automatic Extraction of Synonymy Information:
An Extended Abstract

A Kumaran, Ranbeer Makin, Vijay Pattisapu and Shaik Emran Sharif
Multilingual Systems Research, Microsoft Research India
Bangalore, India

Gary Kacmarcik and Lucy Vanderwende
Natural Language Processing, Microsoft Research
Redmond, Washington, USA
{kumarana, t-ranbm, t-vijpat, t-shaik, garykac, lucyv}@microsoft.com

Abstract Automatic extraction of semantic information, if successful, offers to languages with little or poor resources, the prospects of creating ontological resources inexpensively, thus providing support for commonsense reasoning applications in those languages. In this paper we explore the automatic extraction of synonymy information from large corpora using two complementary techniques: a generic broad-coverage parser for generation of bits of semantic information, and their synthesis into sets of synonyms using sense-disambiguation. To validate the quality of the synonymy information thus extracted, we experiment with English, where appropriate semantic resources are already available. We cull synonymy information from a large corpus and compare it against synonymy information available in several standard sources. We present the results of our methodology, both quantitatively and qualitatively; we show that good quality synonymy information may be extracted automatically from large corpora using the proposed methodology.

Keywords: Automatic synonymy extraction, Linguistic ontological resources, Word-sense Disambiguation, Application for Latent Semantic Analysis.

1 Introduction

Automatic extraction of semantic information, if successful, will prove to be invaluable for languages with little or poor resources in the way of dictionaries, thesauri, etc. It opens up an unprecedented level of access to obscure and underrepresented languages by enabling such projects as automated compilation of lexica, content organization, and multilingual information retrieval. In this project, we explore the automatic construction of synonymy information from large corpora, using two complementary techniques: a generic broad-coverage parsing for generation of bits of semantic information and their synthesis into sets of synonyms based on sense-disambiguation. To validate the quality of the synonymy information extracted by our methodology, we experiment first with English, where appropriate semantic resources are already available as reference. We cull synonymy information from a large corpus, and compare it against the synonymy information available in multiple sources, specifically, the *Oxford English Dictionary* [Simpson and Weiner 1989] and WordNet [Fellbaum 1998].

We first present a "naïve" approach, where we assemble sets of synonyms under the assumption of transitive synonymy. While, the quantitative and qualitative analysis of synonym sets thus constructed present an endemic problem of *semantic drift*, we present a solution methodology based on sense disambiguation to synthesize better quality synonym sets. Finally, we present some quantitative and qualitative evaluations of the results of our refined approach, including, *inter alia*, comparisons with our naïve approach and WordNet data, as well as discussion of possibilities for this technique.

2 Automatic Synonym Extraction from Large Corpora

In this section, we provide a brief description of the two large resources that we used in our experimentation, i.e., WordNet and MindNet.

WordNet The Princeton WordNet [Fellbaum 1998] is a manually constructed lexical database organized by word meanings (as opposed to word forms, as in a dictionary). A part of WordNet, namely, the noun synonyms resemble a thesaurus. Its hierarchical semantic structure describes hypernymy/hyponymy, holonymy/meronymy, and synonymy/antonymy between words. Different word senses are addressed by writing multiple, enumerated lexical entries (they are effectively treated as if they were different words). WordNet is being used as a lexical knowledge base in a wide variety of information retrieval (IR) applications. Since WordNet is handcrafted, it is thorough, complete, expensive and unique. It is thorough and complete because it has been written by professional lexicographers; specifically, [Fellbaum 1998] states that "in terms of coverage, WordNet's goals differ little from those of a good standard college-level dictionary". It is expensive, having taken decades to compile for English alone. WordNets for other languages have been and are being compiled [Global WordNet], but are available primarily in Western European [Euro WordNet] languages. Given the time and resources needed to develop WordNet in a language, it may be a daunting task for most languages of the world, which are constrained by economic resources, market potential, or linguistic expertise.

MindNet MindNet is an automatic ontology extraction system for unrestricted text in English [Vanderwende et al. 2005] [Richardson et al. 1998] that has also been successfully adapted to Japanese [Suzuki et al. 2005] as well. MindNet builds a logical form graph for each sentence using a broad-coverage parser, and extracts semantic relationships among words in that sentence. Such extracted knowledge is accumulated in MindNet, from which all semantic relationships between two words may be explored explicitly through an explorer interface[1]. The corpora we use for extracting semantic information are two machine-readable dictionaries (MRDs), *The American Heritage Dictionary*, 3rd ed. (AHD) and *Longman's Dictionary of Contemporary English* (LDOCE). Although MindNet can be used with any corpus, we use MRDs in order to produce maximal output with which to construct inferences.

[1] An online explorer of dictionary-based MindNet is available at http://research.microsoft.com/mnex/.

For the same of discussion in the following sections, it is important to emphasize the following two caveats: First, the MindNet extract relationships are between words, where as the WordNet is organized by the word senses. Second, for extracting synonymy information, it has been shown that simpler pattern matching techniques may perform well, in [Chodorow et al. 1985] and [Hearst 1992]. We use a broad-coverage parser, due to its ready availability and our goal of ultimately extracting all types of relationships.

3 Naïve Approach

First we compiled all the synonymy relationships MindNet extracted from the MRDs. This compilation consists of expressions of the form "A *syn* B", essentially encoding the fact that A and B are synonymous in some context. From this we synthesized a set of synsets, wherein if A *syn* B and B *syn* C were found in the extracted expressions, then A, B, and C are put into the same synset (i.e., it is inferred that A *syn* C). The naïve approach is thus characterized by transitive synonymy—any set of nodes connected transitively to each other are grouped into the same synset. In addition to *syn* relationships, we incorporated nodes from the hypernymy/hyponymy relations output of MindNet, in order to cover those WordNet leaf synsets that are primarily singletons. In the following two sections we analyze the quality of such synthesis of synsets.

Quantitative Evaluation of the Naïve Approach The naïve approach, working on the *syn* and *hyp* relationships extracted from AHD and LDOCE, produced 49,693 synsets. Table 1 shows a quantitative comparison of synsets formed by MindNet, with those of WordNet.

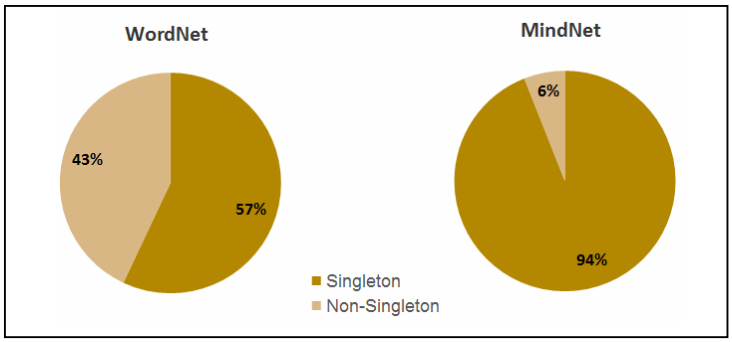
Table 1 Comparison of WordNet and MindNet synsets

Characteristics	WordNet	MindNet
Total Words	117,097	63,230
Total Synsets	81,426	49,693
Avg Words/Synset	1.78	1.08
Avg Synsets/Word	1.24	1.00

We observe that we have only a little more than half as many words and synsets as WordNet, possibly due to the limited extent of corpora that was analyzed by MindNet, resulting in far less number of words for which *syn* information is extracted. We believe that extracting from larger and more diverse corpora might alleviate this relative shortcoming. We see that the synonymy relationships are markedly richer in WordNet, as evidenced by higher averages of words in synset and vice-versa. This is an unsolvable shortcoming of our naïve approach, as a given word could be a part of only one synset, whereas in WordNet, a polysemous word is common to several synsets. Hence, our synthesis must be enhanced to account for polysemous words (which is addressed in the next section). Our subsequent analysis in this section focuses on the quality of the synsets thus extracted by our naïve method, and not on quantity.

Qualitative Evaluation of the Naïve Approach We first analyzed the distribution of sizes of the extracted synsets; as shown in Figure 1, we find that the majority (94%) of the synsets were singletons, produced primarily by the *hyp* relationships, for which there were no corresponding *syn* relationships available. Comparatively, 57% of the WordNet synsets are singletons.

Figure 1 Comparison of WordNet and MindNet synset sizes



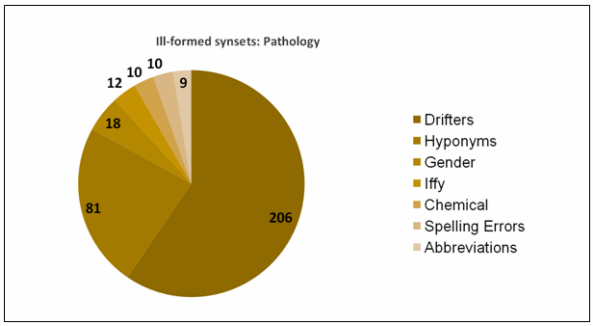
The disparity in proportion of singleton synsets between the two can be due to a variety of reasons. Part of the explanation is the obvious fact that automatic extraction of lexical information will underperforms manual construction of it. Another is that WordNet covers many more words than MindNet's source MRD's. Since the singletons synsets are good, by definition, we examined for quality the remaining 6% (2,882) non-singleton sets, in the subsequent analysis.

We manually inspected the output synsets of the naïve approach against the *Oxford English Dictionary* (OED) and the two source MRD's. Checking against the OED, an independent source, a) prevents artificially high results from using the same corpus as both an input corpus and output test [Schütze 1998], and b) adds insight into the variability of dictionaries, as the disjunction between the correlations of the synset output of OED vs. (AHD + LDOCE) is quantitatively and qualitatively large. Such disjunctions may reveal an importance of corpus choice in automatic ontology extraction; though we aim to be able to handle unrestricted text, we are also interested in exploring the implication of corpus choices on the output quality. The motivation for manually checking the synset output against the AHD and LDOCE is primarily to evaluate the global and local performance of the logical forms produced by broad-coverage parser.

For this manual qualitative analysis, we distinguish between well-formed or ill-formed synsets, which refer only to the quality of the synsets, and not to the quality of MindNet data. First, we found about 87% (2,517) were well-formed, and the rest were ill-formed. Our criteria for whether a synset is well- or ill-formed is an approximation of lexicographers' consensus via manually checking the output against a variety of lexical

resources: OED, AHD, LDOCE, and WordNet. We analyzed manually all the ill-formed synsets and the Figure 2 gives a classification of these synsets.

Figure 2 Pathology of MindNet synthesized synsets



The *drifters* form the majority (206 synsets) of the ill-formed synsets, that is, synsets like {*board*, *committee*, *plank*} that spread across more than one consistent semantic space. In the above example the ill-formed synset, {*board*, *committee*, *plank*} contains two different semantic spaces, namely, {*board*, *committee*} and {*board*, *plank*}. If a synset contains at least one pair of words that are not synonymous, but included due to transitive synonymy, then that synset is classed as a drifter. In our naïve extraction, one pathological case of the drifter synset had nearly 9,500 entries.

Some of the structural patterns encoded in the MindNet extraction algorithm appear to fail in specific cases or in specific syntactic constructions; as evidenced by the accuracy of extracting *syn* and *hyp* relations. Nouns, for example, that appear in conjoined phrases and that can be parsed as either a noun or an adjective, parse preferentially as nouns, even when an adjective parse is called for. The result is that the definition of a dictionary entry as, "differential or integral calculus," incorrectly yields *differential syn integral calculus* by pivoting on the word being defined. We also found other cases of extraction errors found from our analysis. *Chemical* (10) indicates a rather idiosyncratic problem of how the parser handles parentheses. If, for example, the parser encounters the entry for “lead arsenate,” whose empirical formula is $Pb_3(AsO_4)_2$, the resultant synset is {Pb_3, *lead_arsenate*, Cu_3, CO_3, *azurite*, AsO_4, *erythrite*}. This clearly is a result of a wrong parse due to special symbols in chemical formulae, and needs to be corrected. We intend to implement modifications to the MindNet extraction algorithm to increase the extraction accuracy. *Hyponyms* (81) in this chart denotes synsets consisting of hypernyms and their hyponyms. *Gender* (18) denotes gender antonyms like {*actor*, *actress*}; though these pairs fail Leibniz’s Substitution Principle, certain dictionaries’ entries support their synonymy, and at any rate it is rather spurious to worry about these. *Iffy* (12) contains near-synonyms whose validity or invalidity is hard to assert. An example of an “abbreviation” (10) error is the synset {*nm*, *nanometer*, *nuclear_magneton*}. *Spelling* (9) errors are all due to typographic errors in the corpus.

Clearly, drifters are a major problem synthesis of correct synonymy information, and it is clear that the primary reason for their inclusion is the lack of disambiguation between the senses of a word, as MindNet output consists of only words and not word senses. So, the next part of our research focussed on synthesizing the synsets with sense-disambiguation.

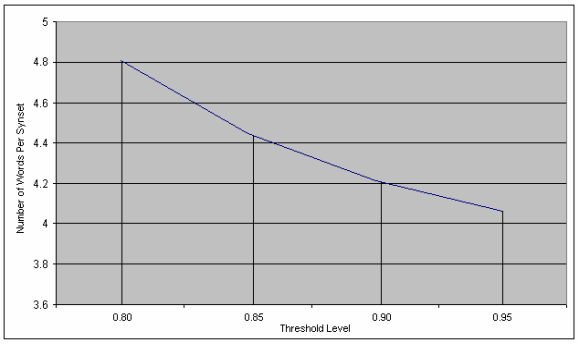
4 Latent Semantic Analysis

In this section, we focus on the Word-sense Disambiguation (WSD) step and how it can filter extracted synonymies into correct synsets. We do this WSD filtering with Latent Semantic Analysis (LSA), a statistical method of assessing words' semantic contexts [Landauer et al. 1993] [Bellegarda 2000]. First, we construct a word-by-document matrix for a large corpus. Next, because of this matrix's sparseness, we extract its principal vectors via singular value decomposition. Finally, we use this information to test the putative synonym pairs provided by MindNet: if the cosine similarity of their vectors is greater than or equal to a threshold, then they are joined into a synset. We hypothesize that the word neighborhood of *plank* will differ sufficiently from the word neighborhood of *committee* so that LSA can thereby "read" two senses of the word *board*. These two approaches—broad-coverage parsing and LSA—are complementary modules in that the former's syntactic approach is blind to parasentential patterns, whereas the latter's "bag-of-words" approach is largely blind to intrasentential patterns. In this experiment, we ran the extraction-side on the two machine-readable dictionaries already mentioned, AHD and LDOCE. We performed LSA on the Brown Corpus[2] [Kucera et al. 1967] to extract a 15-dimensional words space for computing similarity. We note that the number of non-singleton synsets went from nearly 2,800 to nearly 17,000, indicating that a number of large runaway synsets were broken into smaller synsets.

Quantitative Evaluation of LSA-enabled Synthesis We used cosine similarity measure between two words to distinguish their senses, and we used threshold values between 0.8 and 0.95 to empirically study the impact of the threshold on the formation of good synsets. A threshold of 1.0 yields a degenerate solution of cleaving every synset into singletons, and hence was not considered for analysis. While a lower threshold left most good synsets intact, a higher threshold disassociated runaway and loose synsets, creating smaller units, though possibly cleaving even some of the good ones. The words covered in these synsets are exactly same as that presented earlier in the naïve approach, but they are presumably organized differently. We observe, in Figure 3, that the average words per synset (of non-singleton synsets) decreases with the threshold parameter, indicating that the synsets are getting smaller and presumably tighter (an analysis to verify this is provided in a later section).

[2] We tried Encarta in place of Brown corpus, but the results were not good; we hypothesize that the quality could be due to the highly variable document length, improper tuning of LSA parameters [Nakov et al., 2001], or the homogeneous lexical register.

Figure 3 Average sizes of MindNet+LSA synthesized synsets



Also, we compared these synsets with WordNet synsets, purely based on the member words of the synsets, and the results are shown in Table 2. We observe that the number of identical synsets between WordNet and MindNet increases, indicating that the LSA analysis help in building semantically tighter synsets. We also note the two positive trends that the number of synthesized synsets that are subsets of WordNet synsets (thus, well-formed) increase, where as those that contain WordNet synsets (thus, possibly ill-formed) decrease.

Table 2 Comparison of MindNet+LSA synthesized synsets with WordNet synsets

THRESHOLD	0.8	0.85	0.9	0.95
Total Synsets	16,882	17,061	17,581	18,250
Identical to WordNet Synset	1,317	1,361	1,419	1,497
Subset of WordNet Synset	3,281	3,442	3,702	3,944
Superset of WordNet Synset	478	294	156	46

Qualitative Evaluation of LSA-enabled Synthesis Next, we manually examined the synsets synthesized to ascertain the quality: first, we inspected the pre-LSA synset output, tagging synsets with inference-side errors (specifically, drifters) as bad. We then looked at subsets of the good and bad synsets post-LSA (viz., the cleaved synsets). Of the good synsets, 68% remained untouched by the LSA step (i.e., perfect overlap of pre- with post-LSA), while 32% got cleaved (27% of pre-LSA are supersets of their post-LSA counterparts; 5% are partial intersects). Of the bad synsets, meanwhile, everything was cleaved (0% perfect overlap of pre- and post-LSA; 95% of the pre-LSA synsets are supersets of their post-LSA counterparts; 5% are partial intersects), with most of them moving to "good" category.

Next, we selected a random 10% sample of synsets that were not well-formed in the naïve approach, and examined all the synsets in the new synthesis, containing any words that are part of the selected set. The new synsets were classified as *well-formed*, *ill-formed* and *iffy*, as done in the naïve approach, and the results presented in Table 3. We did not

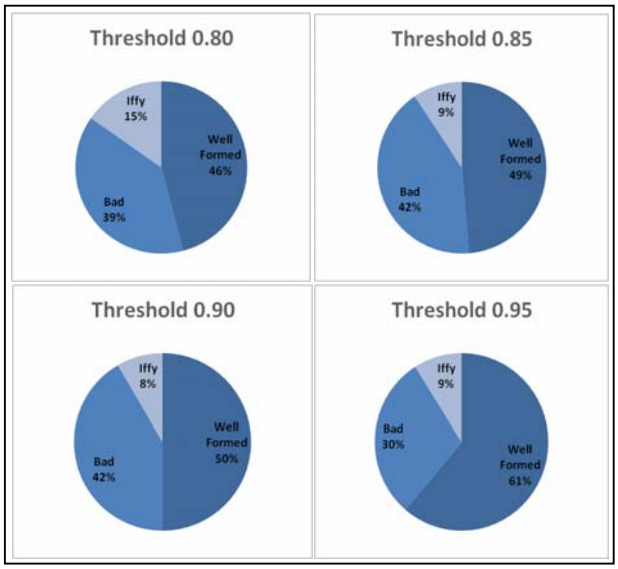
examine any words from the well-formed synsets from the naïve approach, since any synset from naïve approach can only break into smaller pieces, and any subset of a well-formed synset will remain well-formed.

Table 3 Qualitative analysis of MindNet+LSA synthesized synsets

THRESHOLD	0.8	0.85	0.9	0.95
Good Synsets	50%	54%	49%	59%
Bad Synsets	37%	38%	43%	32%
Iffy Synsets	13%	8%	8%	8%

We see that as the similarity threshold increases, the percentage of good synsets increased (as expected, as the synsets get smaller, and possibly, tighter). The growth in the good synsets was mainly due to the cleavage of the drifters from the pre-LSA synsets. The fraction of synsets that were iffy remains the same, indicating that their existence may be due to the extraction side errors.

Figure 4 Word coverage (by classification) in MindNet+LSA synthesized synsets



In addition, as shown in Figure 4, the words are also classified into one of the three categories, in line with that of the synsets. As expected, we see most words that are from bad synsets move into good synset category, with increasing threshold. Overall, we find that nearly half of the not well-formed synsets synthesized by naïve approach could be cleaved into smaller well-formed synsets, showing good promise in extraction of synonymy information using our methodology.

5 Conclusions and Future Work

In this paper, we presented an experiment to automatically acquire a lexical knowledge base, specifically of the type of synonymy information represented in WordNet, using two complementary techniques – a broad-coverage parser for gleaning semantic information from a large corpus and a word-sense disambiguation methodology to synthesize the synsets. To validate our methodology, we conduct this experiment in English, so that the results may be compared directly with WordNet. We use the MindNet system for extracting synonymy information from a set of machine readable dictionaries, specifically the AHD and LDOCE, and construct synonymy using a naïve transitive closure approach. While this approach produced reasonable synsets, we observe the primary shortcoming that a large fraction of the synsets are *drifters*, that is, those that accumulate large unrelated collection of words, due to the polysemous nature of words and the lack of sense disambiguation used in synset construction. Subsequently, we used the result of Latent Semantic Analysis on a large corpus, and use the resulting basis for adding senses of a given word during the synthesis process. A manual analysis indicates that the quality of the resulting synsets improves significantly. Though our proposed methodology did not produce perfect synsets, it shows promise in automatically extracting synsets from natural language text.

The current experiment uses a specific type of natural language text, namely, machine readable dictionaries, but this approach is not limited to dictionaries as many others have demonstrated algorithms to identify definitional text in freely occurring natural text [Saggion and Gaizauskas 2004] and [Joho and Sanderson 2000]. The current experiment also takes its input from Syn and Hyp relations extracted by MindNet using a broad-coverage parser. Naturally, we cannot make the assumption that a parser exists for the language for which we seek to create a WordNet resource, where we can only expect little or no resources. However, other studies have shown that the accuracy for acquiring hypernymy and synonymy using simple string patterns can be as high as 86% for dictionary text [Chodorow et al. 1985], and it is likely that the accuracy will be similarly high for the acquisition from text classified as definitional, using patterns such as described in [Hearst 1992]. We used the synonyms provided by MindNet not to demonstrate that a broad-coverage parser was required, but rather to demonstrate the feasibility of combining automatically extracted synonyms with LSA to produce a lexical knowledge base similar in quality to WordNet. What remains to be seen is the size of the knowledge base we might extract in this manner for a language that might have a smaller body of available text to draw from than languages already studied. However, we anticipate that the knowledge base created can act as a seed for subsequent extensions, such as suggested by [Roark and Charniak 2004] and [Snow et al. 2006]. In combination, these methods will pave the way for unprecedented levels of access to the under-represented languages of the world.

Acknowledgements We thank Jagadeesh Jagarlamudi for his insightful comments and Shravan Kumar for his help in experimentation.

References

1. Bellegarda, J. R. 2000. Exploiting Latent Semantic Information in Statistical Language Modeling. *Proceedings of the IEEE,* Vol. 88, No. 8.
2. Euro WordNet. *http://www.illc.uva.nl/EuroWordNet.*
3. Fellbaum, C., Ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, London.
4. Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter L. A. and Lochbaum, K. E. 1988. *Proceedings of the 11th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*.
5. The Global WordNet Association. *http://www.globalwordnet.org*.
6. Hearst, M. A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics.*
7. Joho, H. and M. Sanderson. 2000. Retrieving descriptive phrases from large amounts of free text. *Proceedings of CIKM*, pages 180-186.
8. Kucera, H. and Francis, W. N. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence RI.
9. Landauer, T. K., Foltz, P. W., and Laham, D. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes 25*: 259-284.
10. Nakov, Preslav, Popova, A. and Mateev, P. 2001. Weight Functions Impact on LSA Performance. *Recent Advances in NLP 2001*.
11. *Oxford English Dictionary.* 2nd ed. 1989 (ed. J. A. Simpson and E. S. C. Weiner), Additions 1993-7 (ed. John Simpson and Edmund Weiner; Michael Proffitt), and 3rd ed. (in progress) Mar. 2000-(ed. John Simpson). Oxford University Press, Oxford, UK.
12. Resnik, P. and Yarowsky, D. 2000. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Natural Language Engineering 5* (3): 113-133.
13. Richardson, S. D., Dolan, W. B., and Vanderwende, L. 1998. MindNet: acquiring and structuring semantic information from text. *Proceedings of the COLING '98*.
14. Roark, B. and Charniak, E. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.
15. Saggion, H. and Gaizauskas, R. 2004. Mining on-line sources for definition knowledge. *Proceedings of the 17th International FLAIRS Conference*.
16. Schütze, H. 1998. Automatic word sense discrimination. *Computational Linguistics*, Vol 24, Issue 1.
17. Snow, R., Jurafsky, D., and Ng, A. 2006. Semantic taxonomy induction from heterogeneous evidence. *Proceedings of COLING/ACL '06*.
18. Suzuki, H., Kacmarcik, G., Vanderwende, L. and Menezes, A. 2005. MindNet / mnex: 意味関係データベースの自動構築と 解析のためのツール (Mindnet and mnex: An Environment for Exploring Semantic Space). 言語処理学会第11回全国大会論文集 *(Proceedings of the 11th Annual Meeting of the Society of Natural Language Processing*).
19. Vanderwende, L. 1995. Ambiguity in the Acquisition of Lexical Information. *AAAI Spring Symposium Series*, No. 95/01, 174-179.
20. Vanderwende, L., Kacmarcik, G., Suzuki, H. and Menezes, A. 2005. MindNet: An Automatically-Created Lexical Resource. *Proc. of HLT/EMNLP 2005 Interactive Demos*.