A Kumaran, Ranbeer Makin, Vijay Pattisapu, Shaik Sharif, Lucy Vanderwende

Evaluating the Quality of Automatically Extracted Synonymy Information

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 common-sense 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 automatic 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, that indicate good quality synonymy information may be extracted automatically from large corpora using the proposed methodology.

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 under-represented 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 automatic sense-disambiguation methodologies. 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 (14) and WordNet (4).

We first present a *naive* 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.

2.1 WordNet

The Princeton WordNet (4) 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, resembles 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 hand-crafted, it is thorough, expensive and unique. It is thorough because it has been created by professional lexicographers; specifically, (4) 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 (6), but are available primarily in Western European (3) languages, and even then in most languages, not as complete as the English WordNet. 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.

2.2 MindNet

MindNet is an automatic ontology extraction system for unrestricted text in English (23) (16) that has also been successfully adapted to Japanese (21) 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

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

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 optimal output for constructing inferences.

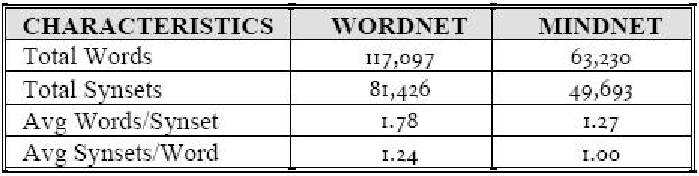
For the sake of discussion in the following sections, it is important to emphasize the following two caveats: First, the relationships extrated by MindNet hold between words, where as WordNet is organized by the word senses. Second, for extracting synonymy information, it has been shown that simpler pattern matching techniques may perform well, in (2) and (7). However, we use a broad-coverage parser, due to its ready availability and due to our goal of ultimately extracting all types of relationships (12).

3 Naive 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.

3.1 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. Figure 1 shows a quantitative comparison of synsets formed by MindNet, with those of WordNet.



CHARACTERISTICS	WORDNET	MINDNET
Total Words	117,097	63,230
Total Synsets	81,426	49,693
Avg Words/Synset	1.78	1.27
Avg Synsets/Word	1.24	1.00

Figure 1: Comparison of WordNet and MindNet Synsets

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 indicated 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.

3.2 Qualitative Evaluation of the Naïve Approach

We first analyzed the distribution of sizes of the extracted synsets; as shown in Figure 2, 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.

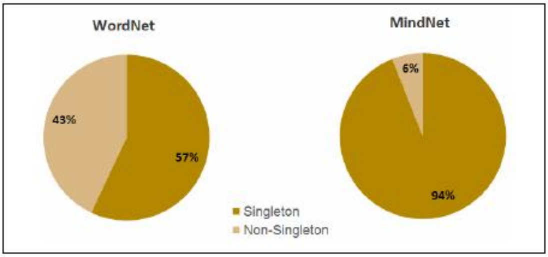


Figure 2: Comparison of WordNet and MindNet Synset Sizes

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 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 machine readable dictionaries. Checking against the OED, an independent source when compared with AHD and LDOCE, has several advantages; first, it prevents artificially high results from using the same corpus as both an input corpus and output test (19). In addition, it adds insight into the variability of dictionaries, as large differences were observed, when the synsets gleaned from AHD + LDOCE were cross-verified with OED, while only minimal variations were expected. Such differences may reveal 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; that is, while the synonymy information captured in the logical forms produced from a single definition in AHD or LDOCE is expected to be correct, we also wish to verify the global consistency of the synonymy information gleaned from logical forms produced by parsing multiple definitions for a set of related words.

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 the MindNet data. 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. Essentially, a synset is classified as well-formed, if each pair of the words from that synset are synonymous, when checked against OED, AHD, LDOCE and WordNet. By this method, we found that about 87% (2, 517) of the extracted synsets were well-formed, and the rest were ill-formed. Next, we analyzed manually all the ill-formed synsets and classified them into different categories, depending on the reasons for their ill-formedness; Figure 3 gives a classification of these ill-formed synsets.

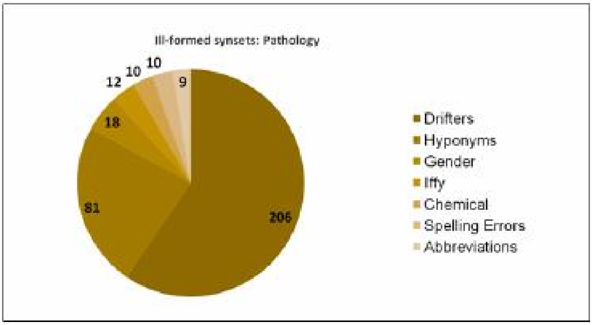


Figure 3: Pathology of MindNet Synthesized Synsets

The drifters form the majority of the ill-formed synsets (206 synsets, constituting of nearly 60% of the ill-formed synsets). Drifters are 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.

In addition to drifters, we find several other classes of problems. Nouns, for example, that appear in coordinate phrases and that can be parsed as either a noun or an adjective, parse preferentially as nouns, resulting often in incorrectly synthesized synsets; for example, the definition of a dictionary entry for "*calculus*" as, "*differential or integral calculus*", incorrectly yields "*differential syn integral calculus*". Such wrong parses resulted in hyponyms and hypernyms grouped in a synset, and accounted for nearly 23% of the errors (tagged as Hyponyms, in Figure 2). Some idiosyncratic, but simpler to correct, parse errors involved chemical names, in which the presence of paran-theses lead to wrong parses; for example, the entry for lead arsenate, whose empirical formula is $Pb_3(AsO_4)_2$, results in a synset { Pb_3, *lead_arsenate*, AsO_4, *azurite*, *erythrite*}, along with that of azurite and erythrite, which share chemical sub-structures with lead arsenate. About 3% of the synsets had similar misclassifications. Gender (5% of the wrongly classified synsets) denotes gender antonyms like { *actor, actress* }, but were classified together, perhaps because they were provided as examples for a hypernym, say, an artist. Though these pairs fail Leibniz's Substitution Principle, certain dictionaries' entries support their synonymy. Iffy (3% of the misclassified synsets) contains near-synonyms whose validity or invalidity is hard to assert. An example of an "abbreviation" (3% of the misclassified synsets) error are synsets such as, { *nm, nanometer, nuclear_magneton* }, where the same abbreviation for two different entity played a role in all of them getting clubbed together. Spelling (3% of the misclassified synsets) 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 (11) (1). First, we construct a word-by-document matrix for a large text 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 Latent Semantic Analysis – 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. In the first set of experiments, we used that any two co-occurring words in the same document are considered associated semantically. In the subsequent analysis, we tighten this assumption, by considering only a window of n words, to form semantic associations.

5 Quantitative Evaluation of Synthesized Synsets

We performed LSA on the Brown Corpus (9) to extract a 15-dimensional words space for computing similarity. The Brown corpus consists of about 1 Million words, $40,897$ unique words distributed among 500 documents. The average words per document is about $2,000$, indicating fairly large documents.

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 the same as those presented earlier in the naïve approach, but they are, understandably, organized differently.

First, 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. We observe, in Figure 4, 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).

We compared these synsets with WordNet synsets, whether the synthesized synset is identical, superset or subset of WordNet synset, purely based on the words of the synsets, and the results are shown in Figure 5.

We observe that the number of identical synsets between WordNet and MindNet increases, indicating that the LSA analysis help in building semantically tighter synsets.

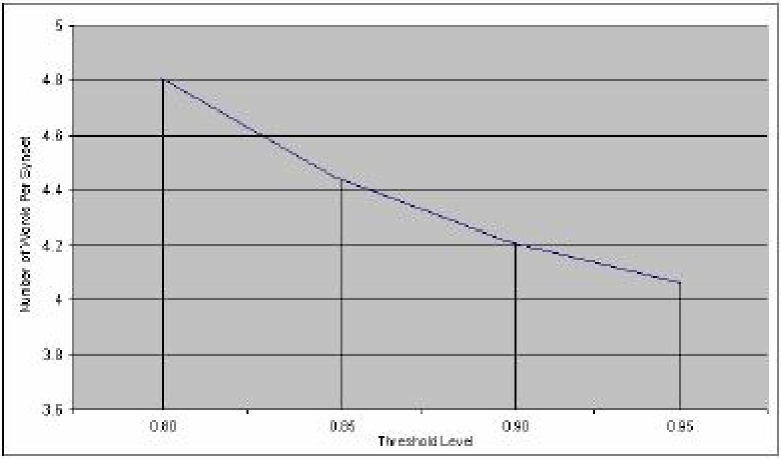
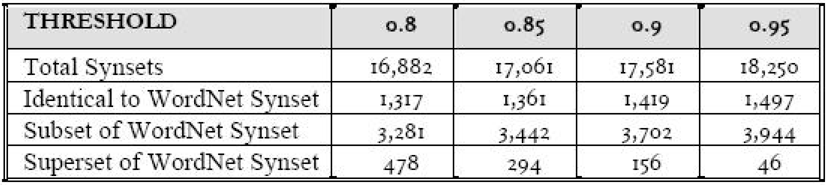


Figure 4: Average Sizes of Mindnet and LSA Synthesized Synsets

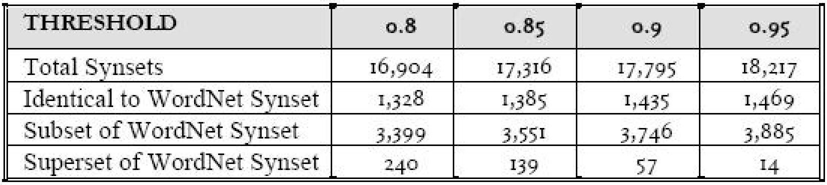


THRESHOLD	0.8	0.85	0.9	0.95
Total Synsets	16,882	17,061	17,581	18,250
Identical to WordNet Synset	1,317	1,361	1,419	1,497
Subset of WordNet Synset	3,281	3,442	3,702	3,944
Superset of WordNet Synset	478	294	156	46

Figure 5: Analysis of Mindnet Synsets Synthesized with WxD LSA on Brown Corpus

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.

The quality of the extracted synsets will directly depend on the quality of the LSA, given that extraction side of synonymy relationships are fixed; hence we experimented with a larger corpus, in order to capture the semantic relationships between different words in a statistically significant way. Hence we chose Microsoft Encarta, a corpus with about 17 Million words, comprising of 173, 807 unique words distributed among 42, 153 documents, to see if the resulting synsets are tighter, when compared with that of WordNet synsets. In addition, the Encarta corpus has about 413 words per document, making the documents smaller, and hopefully providing more meaningful associations between words. It should be noted that the words were not stemmed, but used as they occur in the corpus. In comparison, the Brown corpus has about 1 Million words, distributed in about 500 documents. Figure 6 lists the results of this analysis:



THRESHOLD	0.8	0.85	0.9	0.95
Total Synsets	16,904	17,316	17,795	18,217
Identical to WordNet Synset	1,328	1,385	1,435	1,469
Subset of WordNet Synset	3,399	3,551	3,746	3,885
Superset of WordNet Synset	240	139	57	14

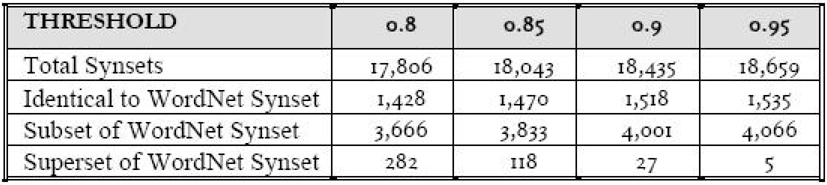
Figure 6: Analysis of Mindnet Synsets Synthesized with WxD LSA on Encarta Corpus

Comparing the results between Figure 5 and Figure 6, we can first observe that the same pattern of variation of the parameters with the threshold. We observe that the number of extracted synsets that are identical to the WordNet synsets is more, as well as those that are subsets of WordNet synsets. The extracted synsets that are superset of WordNet synsets decrease with the threshold. The figures taken together indicate that a small improvement in quality may be achieved by using a larger corpus for LSA analysis, in line with our expectations

5.1 Word-context provided by a Window

In the subsequent set of experiments, we used a word-proximity to measure and quantify the context of a word, as intuitively, such a behaviour might be more meaningful than assuming every word in a document provides the semantic context to a given word. In this procedure, a window representing a span of n words is passed over the corpus being used for LSA analysis, and the window is assumed to provide the context of a word to capture its semantics. The width of the window can be set, but we assumed the size to be 11 (providing 5 words on either side) as the context of a given word. In addition, we assumed a weighting parameter for a context word that is inversely proportional to the number of words separating the context word from the word under consideration, in a given window. Such weights provides a strength for an association between several co-occurring words. Such a window-based context measure provides a co-occurrence matrix that has, as axes, the entire vocabulary found in the corpus. Each cell of the matrix represents the co-occurrence counts for every word pair, weighted appropriately depending on the intervening words. Please note that the word pair, in our discussion, is sensitive to the direction of association; the associations "$x \ldots y$" and the associations "$y \ldots x$" are captured in different cells of the co-occurrence matrix.

We applied the above mentioned methodology on Brown corpus (of about 1 Million words), and by using a window size of 10 words (5 to either sides of the target word), a word by word co-occurrence matrix was created. Such matrix was used for the LSA analysis (in a very similar manner as explained earlier), with the dimensions reduced to 15, using Singular Value Decomposition. Once created, word-by-word cosine similarity measures were used to disambiguate between words. The results of the above experimentation are shown in Figure 7.



THRESHOLD	0.8	0.85	0.9	0.95
Total Synsets	17,806	18,043	18,435	18,659
Identical to WordNet Synset	1,428	1,470	1,518	1,535
Subset of WordNet Synset	3,666	3,833	4,001	4,066
Superset of WordNet Synset	282	118	27	5

Figure 7: Analysis of Mindnet Synsets Synthesized with WxW LSA on Brown Corpus

Comparing the results of word-by-document analysis (as given in Figure 5) and the results by word-by-word analysis (as given in Figure 7), we notice that the same pattern of improvement of quality in all parameters of quantita-tive evaluation. However, there is a significant improvement in quality of the synthesized synsets, even more than that are synthesized by the Encarta corpus. One may conclude that the word-by-word context captures the semantic associations (for the same corpus), and is even more effective than just using larger corpus for analysis[2].

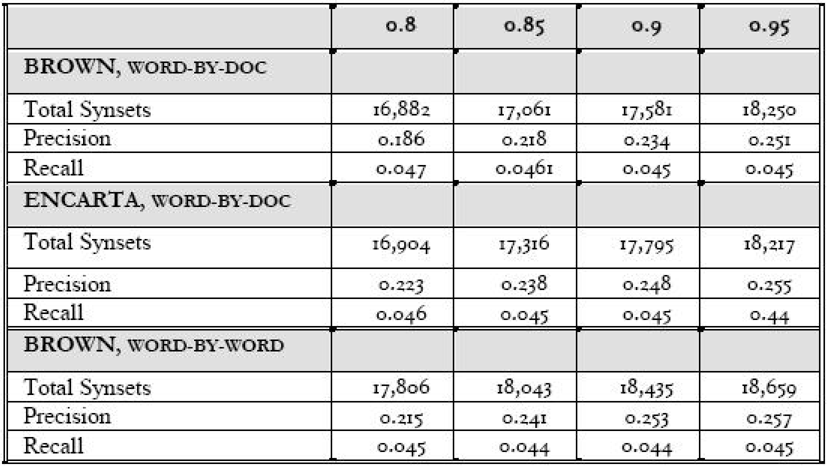
5.2 Verification against WordNet Synsets

While the above verification methodology compared the extracted synsets as a whole, against those of WordNet, we used a second methodology to examine how well the constituents (that is, constituent words) of the extracted synsets measured against the reference synsets, namely the WordNet synsets. While we do present a third strategy (in Section 6) that examines linguistically the quality of the synsets, such a methodology is too expensive, time-wise and resource-wise, to be pursued for the entire set of extracted synsets. In this section, we present our strategy to compare quality of extracted synsets against WordNet synsets in a quantitative manner.

The reference synsets that we use for the evaluation are from WordNet, and we take the hand-crafted WordNet synsets as the gold standard (that is, we do not question the correctness or completeness of the WordNet synsets). In this methodology, we only measure how well our synthesis strategy was able to cover the WordNet synsets.

[2] We were unable to run the word-by-word analysis for ENCARTA corpus, as the resulting size of the word-by-word matrix (with about 173,000 rows and columns each) was too large for our mathematical analysis system to handle.

First we define two metrics, precision and recall, for our extracted synsets against the reference synsets, and subsequently present the two metrics for each of the above synthesis methodology. A word-pair is defined to be a doubleton from a given synset. Hence, given a synset { $s_1, s_2, \cdots s_n$ }, there are n^2 word-pairs in it. Note that though a synset is a set of words, there could be multiple synsets that are associated with a word, corresponding to different senses of the word. Given the above, the precision metric of an extraction of synsets is defined (along the lines of IR systems) as the ratio of the common word-pairs between the extracted synsets and the WordNet synsets, to the total number of word-pairs in the extracted synsets. Similarly, the recall of the synthesized synsets is computed as the ratio between the common word-pairs between the extracted synsets and the WordNet synsets and the total number of word-pairs in the WordNet synsets. In essence, the recall metric indicates the fraction of the information in the WordNet synsets that has been captured in our synthesis, and the precision metric indicates the amount of extraneous information (or noise) present in the extracted synsets.



	0.8	0.85	0.9	0.95
BROWN, WORD-BY-DOC				
Total Synsets	16,882	17,061	17,581	18,250
Precision	0.186	0.218	0.234	0.251
Recall	0.047	0.0461	0.045	0.045
ENCARTA, WORD-BY-DOC				
Total Synsets	16,904	17,316	17,795	18,217
Precision	0.223	0.238	0.248	0.255
Recall	0.046	0.045	0.045	0.44
BROWN, WORD-BY-WORD				
Total Synsets	17,806	18,043	18,435	18,659
Precision	0.215	0.241	0.253	0.257
Recall	0.045	0.044	0.044	0.045

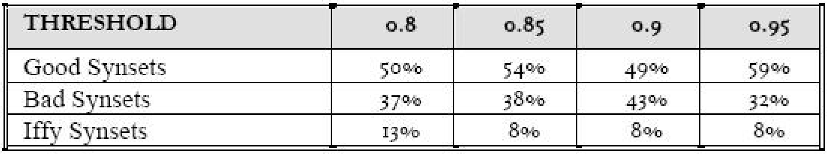
Figure 8: Analysis of Precision and Recall of Synthesized Synsets

Once the metrics were specified as above, the corresponding values of these metrics for each of our extraction methodology may be com-puted automatically. Figure 8 provides the results, from which, we could infer the following: First, the recall metric is very similar in all methodologies; this is to be expected, since the extraction-side is the same for all methodologies. Hence, irrespective of the methodology that is used for synthesis of the synsets, the same words would have been used, and hence recall is bound to be similar. Second, the recall values are small (4.5%), for all extraction methodologies, which may be due to the following reasons: First, our strategy extracted only about 18, 000 synsets (compared with about 80, 000 in WordNet; second, out of the extracted synsets only about 1, 500 are exactly same as the WordNet synsets, and we have nearly three-times as many synsets that are proper subsets of the WordNet synsets (thus yielding less number of word-pairs than the corresponding WordNet synsets) and, third, very few of the synthesized synsets are super-sets of the WordNet synsets. The precision of synthesized synsets sense-disambiguated using larger corpus are markedly better, in-line with our intuition. Further, we see that the disambiguation is better with a context provided by words around a given word, than by the entire document. We are currently, experimenting with different definitions of precision and recall metrics, in order to arrive at intuitive ones.

6 Qualitative Evaluation of Synthesized Synsets

Next, we manually examined the synsets synthesized to ascertain the quality, using the following procedure: 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, every-thing 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 Figure 9.



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%

Figure 9: Classification of Synthesized Synsets

It should be noted that 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. 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.

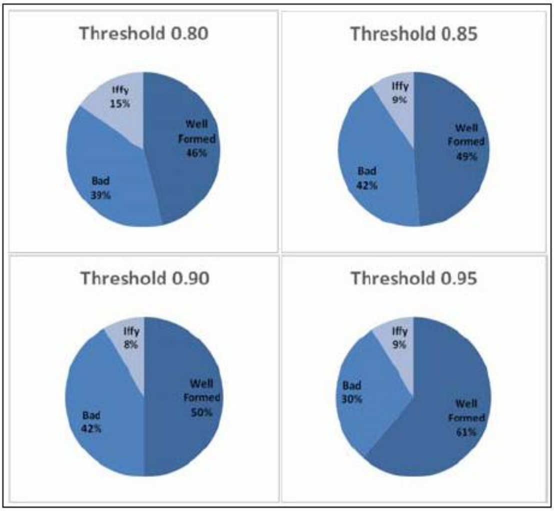


Figure 10: Word Coverage by Classification in the Synthesized Synsets

In addition, as shown in Figure 10, 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.

7 Conclusions and Future Work

In this paper, we presented an experiment to automatically acquire a lexical knowledge base of the same type as 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 used 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 used 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 promising results 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, as in, (18) and (8). 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 (2), 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 (7). 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 shown 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 (17) and (20). 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 discussions, and Gary Kacmarcik for his valuable help in experimentations.*

8 References

References

Bellegarda, J. R. Exploiting Latent Semantic Information in Statistical Language Modeling. *Proceedings of the IEEE, Vol. 88, No. 8*, 2000.

Chodorow, M., Byrd, R. J. and Heidorn, G. E. Extracting Semantic Hierarchies from a Large On-Line Dictionary. *Proceedings of the ACL*, 1985.

Euro WordNet. *http://www.illc.uva.nl/EuroWordNet.*

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

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter L. A. and Lochbaum, K. E. *Proceedings of the 11th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 1988.

The Global WordNet Association. *http://www.globalwordnet.org*.

Hearst, M. A. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics*, 1992.

Joho, H. and M. Sanderson. Retrieving descriptive phrases from large amounts of free text. *Proceedings of CIKM, pages 180-186*, 2000.

Kucera, H. and Francis, W. N. Computational Analysis of Present-Day American English. *Brown University Press, Providence RI*, 1967.

Kumaran, A., Makin, R., Pattisapu, V., Sharif, S. E., Kacmarchi, G. and Vanderwende, L. Automatic Extraction of Synonymy Information: An Extended Abstract. *Proceedings of the Ontologies in Text Technology Workshop*, 2006.

Landauer, T. K., Foltz, P. W., and Laham, D. Introduction to Latent Semantic Analysis. *Discourse Processes 25: 259-284*, 1998.

Montemagni, S. and Vanderwende, L. Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries. *Proceedings of COLING*, 1992.

Nakov, Preslav, Popova, A. and Mateev, P. Weight Functions Impact on LSA Performance. *Recent Advances in NLP*, 2001.

Oxford English Dictionary. 2nd ed. 1989 (ed. J. A. Simpson and E. S. C. Weiner). *Oxford University Press, Oxford, UK.*

Resnik, P. and Yarowsky, D. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Natural Language Engineering 5 (3): 113-133*, 2000.

Richardson, S. D., Dolan, W. B., and Vanderwende, L. MindNet: acquiring and structuring semantic information from text. *Proceedings of the COLING*, 1998.

Roark, B. and Charniak, E. 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*, 1998.

Saggion, H. and Gaizauskas, R. Mining on-line sources for definition knowledge. *Proceedings of the 17th International FLAIRS Conference*, 2004.

Schütze, H. Automatic word sense discrimination. *Computational Linguistics, Vol 24, Issue 1*, 1998.

Snow, R., Jurafsky, D., and Ng, A. Semantic taxonomy induction from heterogeneous evidence. *Proceedings of COLING/ACL*, 2006.

Suzuki, H., Kacmarcik, G., Vanderwende, L. and Menezes, A. MindNet / mnex: An Environment for Exploring Semantic Space). *Proceedings of the 11th Annual Meeting of the Society of Natural Language Processing*, 2005.

Vanderwende, L. Ambiguity in the Acquisition of Lexical Information. *AAAI Spring Symposium Series, No. 95/01, 174-179*, 1995.

Vanderwende, L., Kacmarcik, G., Suzuki, H. and Menezes, A. MindNet: An Automatically-Created Lexical Resource. *Proc. of HLT/EMNLP*, 2005.