Differentiating Homonymy and Polysemy
in Information Retrieval

Christopher Stokoe
School of Computing and Technology
University of Sunderland
UK
`christopher.stokoe@sunderland.ac.uk`

Abstract

Recent studies into Web retrieval have shown that word sense disambiguation can increase retrieval effectiveness. However, it remains unclear as to the minimum disambiguation accuracy required and the granularity with which one must define word sense in order to maximize these benefits. This study answers these questions using a simulation of the effects of ambiguity on information retrieval. It goes beyond previous studies by differentiating between homonymy and polysemy. Results show that retrieval is more sensitive to polysemy than homonymy and that, when resolving polysemy, accuracy as low as 55% can potentially lead to increased performance.

1 Introduction

Lexical ambiguity refers to words that share the same orthography but have different meanings (word senses). It can be sub-divided into two distinct types, homonymy and polysemy. Homonymy describes when two senses of a given word (or derivation) are distinct. Typically, they are separated by etymology and are therefore entirely unrelated in meaning. One classic example (Kilgarriff, 1992) is '*bat*' as in an airborne mammal (from the Middle English word '*bakke*' meaning flying rodent) vs. '*bat*' as in an instrument used in the game of cricket (from the Celtic for stick or cudgel). There is no underlying relationship between these two meanings which have come about independ-

ently from differing root languages. Alternatively, polysemy describes where two senses of a word are related in that they share membership of a subsuming semantic classification. Consider the word '*mouth*' as in a part of the body vs. '*mouth*' as in the outlet of a river. Both meanings are subsumed by a higher concept (in this case they both describe an opening). Homonymy and polysemy are differentiated in most dictionaries by the major (homonyms) and minor (polysemes) entries for a given word. Where a lexical resource is described in terms of granularity a coarse-grained approach only differentiates between homonymy whereas a fine-grained approach also considers polysemy.

The use of word sense disambiguation in Information Retrieval (IR) has been an active field of study for the past 30 years. Despite several failures (described in Sanderson, 2000) recent studies have begun to show increased retrieval effectiveness, particularly in Web retrieval. However, two key questions remain: (1) to what accuracy must disambiguation be performed in order to show increased retrieval effectiveness and (2) to what level of granularity should disambiguation be performed in order to maximize these gains? This study answers these questions by simulating the impact of ambiguity and its subsequent resolution on retrieval effectiveness.

2 Related Work

The motivation for this research is taken from recent studies (section 2.1) which have demonstrated increased retrieval effectiveness by accounting for word sense. The methodology is derived from previous studies (section 2.2) which model the impact that ambiguity and its subsequent resolution have on IR.

2.1 Accounting for Sense in IR

One of the first studies to show increased retrieval effectiveness through resolving ambiguity was Schütze and Pederson (1995). They used clustering to discriminate between alternate uses of a word. The clusters they produced were apparently fine-grained, although it is not clear if this observation was made with reference to a particular lexical resource. In terms of the accuracy to which they could discriminate meaning, a limited evaluation using a 10 word sample demonstrated accuracy approaching 90%. Results showed that retrieval effectiveness increased when documents were indexed by cluster as opposed to raw terms. Performance further increased when a word in the collection was assigned membership of its three most likely clusters. However, it is not clear if assigning multiple senses leads to coarser granularity or simply reduces the impact of erroneous disambiguation.

Stokoe et al. (2003) showed increased retrieval effectiveness through fine-grained disambiguation where a word occurrence in the collection was assigned one of the sense definitions contained in WordNet. The accuracy of their disambiguation was reported at 62% based on its performance over a large subset of SemCor (a collection of manually disambiguated documents). It remains unclear how accuracy figures produced on different collections can be compared. Stokoe et al. (2003) did not measure the actual performance of their disambiguation when it was applied to the WT10G (the IR collection used in their experiments). This highlights the difficulty involved in quantifying the effects of disambiguation within an IR collection given that the size of modern collections precludes manual disambiguation.

Finally, Kim et al. (2004) showed gains through coarse-grained disambiguation by assigning all nouns in the WT10G collection (section 3) membership to 25 top level semantic categories in WordNet (for more detail about the composition of WordNet see section 4). The motivation behind coarse-grained disambiguation in IR is that higher accuracy is achieved when only differentiating between homonyms. Several authors (Sanderson, 2000; Kim et al., 2004) postulate that fine-grained disambiguation may not offer any benefits over coarse-grained disambiguation which can be performed to a higher level of accuracy.

2.2 The Effects of Ambiguity on IR

The studies described in section 2.1 provide empirical evidence of the benefits of disambiguation. Unfortunately, they do not indicate the minimum accuracy or the optimal level of granularity required in order to bring about these benefits. Perhaps more telling are studies which have attempted to quantify the effects of ambiguity on IR.

Sanderson (1994) used pseudowords to add additional ambiguity to an IR collection. Pseudowords (Gale et al., 1992) are created by joining together randomly selected constituent words to create a unique term that has multiple controlled meanings. Sanderson (1994) offers the example of "*banana/kalashnikov*". This new term features two pseudosenses '*banana*' and '*kalashnikov*' and is used to replace any occurrences of the constituent words in the collection, thus introducing additional ambiguity. In his study, Sanderson experimented with adding ambiguity to the Reuters collection. Results showed that even introducing large amounts of additional ambiguity (size 10 pseudowords - indicating they had 10 constituents) had very little impact on retrieval effectiveness. Furthermore, attempts to resolve this ambiguity with less than 90% accuracy proved extremely detrimental.

Sanderson (1999) acknowledged that pseudowords are unlike real words as the random selection of their constituents ensures that the pseudosenses produced are unlikely to be related, in effect only modeling homonymy. Several studies (Schütze, 1998; Gaustad, 2001) suggest that this failure to model polysemy has a significant impact. Disambiguation algorithms evaluated using pseudowords show much better performance than when subsequently applied to real words. Gonzalo et al. (1998) cite this failure to model related senses in order to explain why their study into the effects of ambiguity showed radically different results to Sanderson (1994). They performed known item retrieval on 256 manually disambiguated documents and showed increased retrieval effectiveness where disambiguation was over 60% accurate. Whilst Sanderson's results no longer fit the empirical data, his pseudoword methodology does allow us to explore the effects of ambiguity without the overhead of manual disambiguation. Gaustad (2001) highlighted that the challenge lies in adapting pseudowords to account for polysemy.

Krovetz (1997) performed the only study to date which has explicitly attempted to differentiate between homonymy and polysemy in IR. Using the Longmans dictionary he grouped related senses based on any overlap that existed between two sense definitions for a given word. His results support the idea that grouping together related senses can increase retrieval effectiveness. However, the study does not contrast the relative merits of this technique against fine-grained approaches, thus highlighting that the question of granularity remains open. Which is the optimal approach? Grouping related senses or attempting to make fine-grained sense distinctions?

3   Experimental Setup

The experiments in this study use the WT10G corpus (Hawking and Craswell, 2002), an IR web test collection consisting of 1.69 million documents. There are two available Query / Relevance Judgments sets each consisting of 50 queries. This study uses the TREC 10 Web Track Ad-Hoc query set (NIST topics 501 – 550). The relevance judgments for these queries were produced using pooling based on the top 100 ranked documents retrieved by each of the systems that participated in the TREC 10 Web Track.

Initially the author produced an index of the WT10G and performed retrieval on this unmodified collection in order to measure baseline retrieval effectiveness. The ranking algorithm was length normalized TF.IDF (Salton and McGill, 1983) which is comparable to the studies in section 2. Next, two modified versions of the collection were produced where additional ambiguity in the form of pseudowords had been added. The first used pseudowords created by selecting constituent pseudosenses which are unrelated, thus introducing additional homonymy. The second used a new method of generating pseudowords that exhibit polysemy (the methodology is described in section 4.1). Contrasting retrieval performance over these three indexes quantifies the relative impact of both homonymy and polysemy on retrieval effectiveness. The final step was to measure the effects of attempting to resolve the additional ambiguity which had been added to the collection. In order to do this, the author simulated disambiguation to varying degrees of accuracy and measured the impact that this had on retrieval effectiveness.

4   Methodology

To date only Nakov and Hearst (2003) have looked into creating more plausible pseudowords. Working with medical abstracts (MEDLINE) and the controlled vocabulary contained in the MESH hierarchy they created pseudosense pairings that are related. By identifying pairs of MESH subject categories which frequently co-occurred and selecting constituents for their pseudowords from these pairings they produced a disambiguation test collection. Their results showed that category based pseudowords provided a more realistic test data set for disambiguation, in that evaluation using them more closely resembled real words. The challenge in this study lay in adapting these ideas for open domain text.

4.1   Pseudoword Generation

This study used WordNet (Miller et al., 1990) to inform the production of pseudowords. WordNet (2.0) is a hierarchical semantic network developed at Princeton University. Concepts in WordNet are represented by synsets and links between synsets represent hypernmy (subsumes) and hyponymy (subsumed) relationships in order to form a hierarchical structure. A unique word sense consists of a lemma and the particular synset in which that lemma occurs. WordNet is a fine-grained lexical resource and polysemy can be derived to varying degrees of granularity by traversing the link structure between synsets (figure 1).
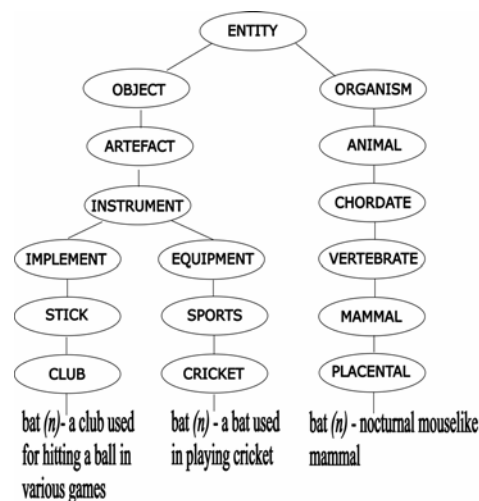

Figure 1. A Subsection of the Noun Hierarchy in WordNet

An important feature of pseudowords is the number of constituents as this controls the amount of additional ambiguity created. A feature of all previous studies is that they generate pseudowords with a uniform number of constituents, e.g. size 2, size 5 or size 10, thus introducing uniform levels of additional ambiguity. It is clear that such an approach does not reflect real words given that they do not exhibit uniform levels of ambiguity. The approach taken in this study was to generate pseudowords where the number of constituents was variable. As each of the pseudowords in this study contain one query word from the IR collection then the number of constituents was linked directly to the number of senses of that word contained in WordNet. This effectively doubles the level of ambiguity expressed by the original query word. If a query word was not contained in WordNet then this was taken to be a proper name and exempted from the process of adding ambiguity. It was felt that to destroy any unambiguous proper names, which might act to anchor a query, would dramatically overstate the effects of ambiguity in terms of the IR simulation. The average size of the pseudowords produced in these experiments was 6.4 pseudosenses.

When producing the traditional pseudoword based collection the only modification to Sanderson's (1994) approach (described in section 2), other than the variable size, involved formalizing his observation that the constituent words were unlikely to be related. Given access to WordNet it was possible to guarantee that this is the case by rejecting constituents which could be linked through its inheritance hierarchy. This ensures that the pseudowords produced only display homonymy.

In order to produce pseudowords that model polysemy it was essential to devise a method for selecting constituents that have the property of relatedness. The approach taken was to deliberately select constituent words that could be linked to a sense of the original query word through WordNet. Thus the additional ambiguity added to the collection models any underlying relatedness expressed by the original senses of the query word. Pseudowords produced in this way will now be referred to as root pseudowords as this reflects that the ambiguity introduced is modeled around one root constituent. Consider the following worked example for the query "How are tornadoes formed?"

After the removal of stopwords we are left with 'tornadoes' and 'formed' each of which is then transformed into a root pseudoword. The first step involves identifying any potential senses of the target word from WordNet. If we consider the word 'tornado' it appears in two synsets:

1. tornado, twister -- (a localized and violently destructive windstorm occurring over land characterized by a funnel-shaped cloud extending toward the ground)

2. crack, tornado -- (a purified and potent form of cocaine that is smoked rather than snorted)

For each sense of the target word the system expands WordNet's inheritance hierarchy to produce a directed graph of its hypernyms. Figure 2 shows an example of this graph for the first sense of the word 'tornado'. In order to ensure a related sense pair the system builds a pool of words which are subsumed by concepts contained in this graph. This is generated by recursively moving up the hierarchy until the pool contains at least one viable candidate. For a candidate to be viable it must meet the following criteria:

1) It must exist in the IR collection.
2) It must not be part of another pseudoword.
3) It can not be linked (through WordNet) to another constituent of the pseudoword.

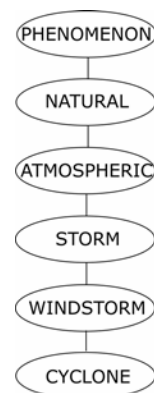The pool for sense 1 of 'tornado' consists of [*hurricane|typhoon*], one of which is selected at random.

Figure 2. A Graph of the Hypernyms for the First Sense of the Word 'tornado'

This process is repeated for each noun and verb sense of the query word. In this example there is one remaining sense of the word '*tornado*' - a slang term used to refer to the drug crack cocaine. For this sense the system produced a pool consisting of [*diacetylemorphine*/*heroin*]. Once all senses of the query word have been expanded the resulting pseudoword, '*tornadoes/hurricane/heroin*', is then used to replace all occurrences of its constituents within the collection. Through this process the system produces pseudowords with pseudosense pairings which have subsuming relationships, e.g. '*tornadoes/hurricane*' are subsumed by the higher category of '*cyclone*' whilst '*tornadoes/heroin*' are subsumed by the higher semantic category of '*hard_drug*'.

4.2   Pseudo-disambiguation

In order to perform pseudo-disambiguation the unmodified collection acts as a gold standard model answer. Through reducing each instance of a pseudoword back to one of its constituent components this models the selection process made by a disambiguation system. Obviously, the correct pseudosense for a given instance is the original word which appeared at that point in the collection. Variable levels of accuracy are introduced using a weighted probability model where the correct pseudosense for a given test instance is seeded with a fixed probability equivalent to the desired accuracy being simulated. When a disambiguation error is simulated one of the incorrect pseudosenses is selected randomly.

5   Results

The first set of results (section 5.1) addresses the question of granularity by quantifying the impact that adding either additional homonymy or polysemy has on retrieval effectiveness. The second set of results (section 5.2) looks at the question of disambiguation accuracy by simulating the impact that varied levels of accuracy have on retrieval effectiveness.

5.1   Homonymy vs. Polysemy

Let us first consider the impact of adding additional homonymy. Figure 3 graphs precision across the 11 standard points of recall for retrieval from both the baseline collection and one where addi-

tional homonymy has been added. Note that the introduction of additional homonymy brings about a small drop in retrieval effectiveness. With regard to the single value measures contained in table 1, this is a decrease of 2.5% in terms of absolute R-Precision (average precision after the total number of known relevant documents in the collection has been retrieved). This is a relative decrease of 14.3%. Similar drops in both precision@10 (precision after the first 10 documents retrieved) and average precision are also seen.

Next let us consider retrieval effectiveness over the root pseudoword collection where additional polysemy has been added (figure 4). Note that the introduction of additional polysemy has a more substantive impact upon retrieval effectiveness. In terms of R-Precision this decrease is 5.3% in absolute terms, a relative decrease of 30% compared to baseline retrieval from the unmodified collection. In addition, an even larger decrease in precision@10 occurs where the introduction of additional polysemy brings about a 7% drop in retrieval effectiveness.

In terms of the relative effects of homonymy and polysemy on retrieval effectiveness then note that adding additional polysemy has over double the impact of adding homonymy. This provides a clear indication that the retrieval process is more substantially affected by polysemy than homonymy.
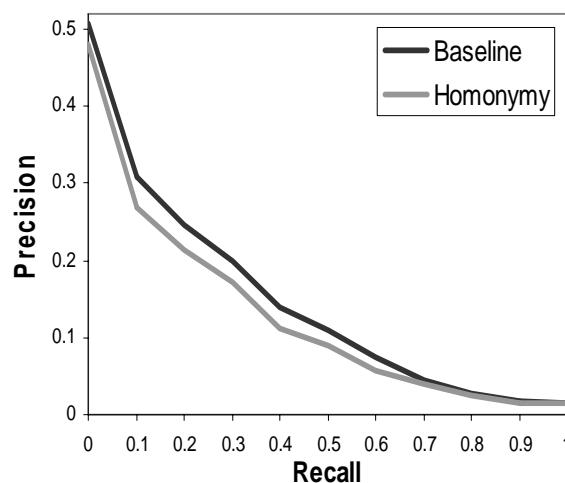

Figure 3. Precision across the 11 Standard Points of Recall for the Baseline and the Collection Containing Additional Homonymy

	R-Precision	Precision @10	Avg. Precision
Baseline	0.1732	0.2583	0.1334
Homonymy	0.1485	0.2208	0.1145
Polysemy	0.1206	0.1875	0.0951

Table 1. R-Precision, Precision@10 and Average Precision for all Three Versions of the Collection
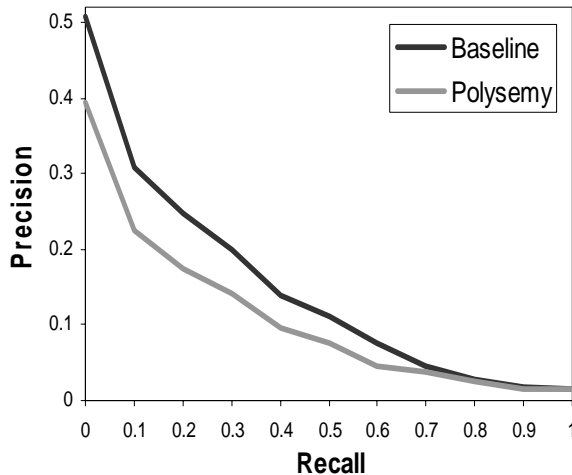

Figure 4. Precision across the 11 Standard Points of Recall for the Baseline and the Collection Containing Additional Polysemy

5.2 The Impact of Disambiguation

We now address the second part of the research question: to what accuracy should disambiguation be performed in order to enhance retrieval effectiveness? Figure 5 plots the impact, in terms of R-Precision, of performing disambiguation to varying degrees of accuracy after additional homonymy has been added to the collection. The dotted line represents the breakeven point, with R-Precision below this line indicating reduced performance as a result of disambiguation. Results show that where additional homonymy has been added to the collection disambiguation accuracy at or above 76% is required in order for disambiguation to be of benefit. Performing disambiguation which is less than 76% accurate leads to lower performance than if the additional homonymy had been left unresolved.

Moving on to consider the root pseudoword collection (figure 6) note that the breakeven point is only 55% where additional polysemy has been

added. Consider that the results in section 5.1 showed that the introduction of additional polysemy had over double the impact of introducing additional homonymy. This is reflected in the relative effects of disambiguation in that the breakeven point is considerably lower for polysemy than homonymy.
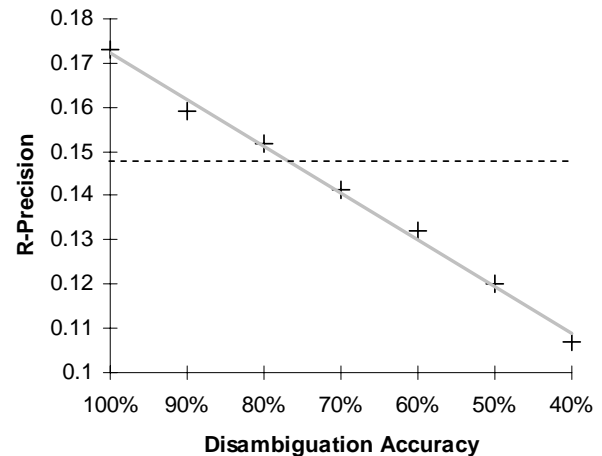

Figure 5. The Impact of Disambiguation on Effectiveness after the Addition of Homonymy

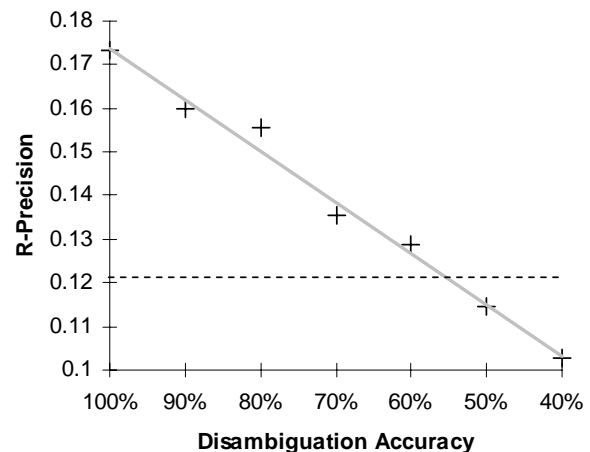**(Note the dashed line is the breakeven point)**

Figure 6. The Impact of Disambiguation on Effectiveness after the Addition of Polysemy

(Note the dashed line is the breakeven point)

6 Discussion

The results in section 5.1 show that retrieval effectiveness is more sensitive to polysemy than homonymy. One explanation for this can be

hypothesized from previous studies (Krovetz and Croft, 1992; Sanderson and Van Rijsbergen, 1999) which highlight the importance of co-occurrence between query words. Where two (or more) words appear together in a query, statistical retrieval inherently performs some element of disambiguation. However, in the case of a word with many closely related senses, co-occurrence between query words may not be sufficient for a given sense to become apparent. This is particularly exasperated in Web retrieval given that the average query length in these experiments was 2.9 words. Clearly, the inherent disambiguation performed by statistical IR techniques is sensitive to polysemy in the same way as systems which explicitly perform disambiguation.

With regard to disambiguation accuracy and IR (section 5.2) these experiments establish that performance gains begin to occur where disambiguation is between 55% and 76%. Where within this range the actual breakeven point lies is dependent on the granularity of the disambiguation and the balance between polysemy and homonymy in a given collection. Consider that coarse-grained disambiguation is frequently advocated on the basis that it can be performed more accurately. Whilst this is undoubtedly true these results suggest that homonymy has to be resolved to a much higher level of accuracy than polysemy in order to be of benefit in IR.

It would seem prudent to consider the results of this study in relation to the state-of-the-art in disambiguation. At Senseval-3 (Mihalcea et al., 2004) the top systems were considered to have reached a ceiling, in terms of performance, at 72% for fine-grained disambiguation and 80% for coarse-grained. When producing the English language test collections the rate of agreement between humans performing manual disambiguation was approximately 74%. This suggests that machine disambiguation has reached levels comparable to the performance of humans. In parallel with this the IR community has begun to report increased retrieval effectiveness through explicitly performing disambiguation to varying levels of granularity.

A final point of discussion is the way in which we simulate disambiguation both in this study and those previously (Sanderson, 1994; Gonzalo et al., 1998). There is growing evidence (Leacock et al., 1998; Agirre and Martinez, 2004) to suggest that simulating uniform rates of accuracy and error

across both words and senses may not reflect the performance of modern disambiguation systems. Supervised approaches are known to exhibit inherent bias that exists in their training data. Examples include Zipf's law (Zipf, 1949) which denotes that a small number of words make up a large percentage of word use and Krovetz and Croft's (1992) observation that one sense of a word accounts for the majority of all use. It would seem logical to presume that supervised systems show their best performance over the most frequent senses of the most frequent words in their training data. Not enough is known about the potential impact of these biases to allow for them to be incorporated into this simulation. Still, it should be noted that Stokoe et al. (2003) utilized frequency statistics in their disambiguator and that a by-product of Schütze and Pederson's (1992) approach was that they eliminated infrequently observed senses. There is supporting evidence from Sanderson and Van Rijsbergen (1999) to suggest that accounting for this frequency bias is in some way advantageous. Therefore, it is worth considering that simulating a uniform accuracy and error rate across all words and senses might actually offer a pessimistic picture of the potential for disambiguation and IR. Whilst this merits further study, the focus of this research was contrasting the relative effects of two types of ambiguity and both models were subject to the same uniform disambiguation.

7 Conclusions

This study has highlighted that retrieval systems are more sensitive to polysemy than homonymy. This leads the author to conclude that making fine-grained sense distinctions can offer increased retrieval effectiveness in addition to any benefits brought about by coarse-grained disambiguation. It also emphasises that although coarse-grained disambiguation can be performed to a higher degree of accuracy, this might not directly translate to increased IR performance compared to fine-grained approaches. This is in contrast to current thinking which suggests that coarse-grained approaches are more likely to bring about retrieval performance because of their increased accuracy.

In terms of disambiguation accuracy and increased retrieval effectiveness, results show potential benefits where accuracy is as low as 55% when dealing with just polysemy and rises to 76% when

dealing with just homonymy. Obviously this study has simulated two extremes (polysemy or homonymy) and the exact point where performance increases will occur is likely to be dependent on the interaction between homonymy and polysemy in a given query.

With regard to simulation a more empirical exploration of the ideas expressed in this work would be desirable. However, the size of modern IR test collections dictates that future studies will need to rely more heavily on simulation. Therefore, until such time that a significant manually disambiguated IR collection exists pseudowords remain an interesting way to explore the effects of ambiguity within a large collection. The challenge lies in producing pseudowords that better model real words.

References

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

Gale, W; Church, K. W; Yarowsky, D. 1992 *Work on Statistical Methods for Word Sense Disambiguation,* in Intelligent Probabilistic Approaches to Natural Language Papers, AAAI Press, Pp. 54 – 60.

Gaustad, T. 2001. *Statistical Corpus-based Word Sense Disambiguation: Pseudowords vs. Real Ambiguous Words,* in Companion Volume to the Proceedings of the 36th annual meeting of the ACL, Toulouse, France. Proceedings of the Student Research Workshop, ACL Press, Pp. 61 – 66.

Gonzalo, J; Verdejo, F; Chugur, I; Cigarran, J. 1998. *Indexing with WordNet Synsets Can Improve Text Retrieval,* in Proceedings of COLING / ACL '98 Workshop on the Usage of WordNet for NLP, Montreal, Canada, ACL Press, Pp. 38 – 44.

Hawking, D and Craswell, N. 2002. *Overview of the TREC-2001 Web Track,* in Proceedings of the 10th Text REtrieval Conference, Gaithersburg, MD. NIST Special Publication 500-250, Pp. 61 – 67.

Kilgarriff, A. 1992. *Polysemy*, Ph.D. Thesis, School of Cognitive and Computing Sciences, University of Sussex, Report CSRP 261

Kim, S; Seo, H; Rim, H. 2004. *Information Retrieval Using Word Senses: Root Sense Tagging Approach,* in Proceedings of the 27th ACM SIGIR, Sheffield, UK. Pp. 258 – 265.

Krovetz, R and Croft, W. B. 1992. *Lexical Ambiguity and Information Retrieval,* ACM Transactions on Information Systems, 10(2), Pp. 115 – 141.

Krovetz, R 1997. *Homonymy and Polysemy in Information Retrieval*, NEC Institute, Princeton.

Leacock, C; Chodorow, M; Miller, G. A. 1998 *Using Corpus Statistics and WordNet Relations for Sense Identification.* Computational Linguistics, 24(1), Pp. 147 – 165.

Mihalcea, R; Chklovski, T; Kilgarriff, A. 2004. *The Senseval-3 English Lexical Sample Task,* in Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, Pp. 25 – 28.

Miller, G; Beckwith, R; Fellbaum, C; Gross, D; Miller, K. 1990. *WordNet: An On-line Lexical Database,* International journal of lexicography, 3(4), Oxford University Press, Pp 235 – 244.

Nakov, P. and Hearst , M. 2003. *Category-based Pseudowords,* in Proceedings of the Human Language Technology Conference (HLT-NAACL 2003), Edmonton. Pp. 67–69.

Salton G and McGill, M. J. 1983. *Introduction to Modern Information Retrieval*, McGraw-Hill.

Sanderson, M. 1994. *Word Sense Disambiguation and Information Retrieval,* in Proceedings of the 17th ACM SIGIR, Dublin, Ireland, ACM Press, Pp. 142 – 151.

Sanderson, M and Van Rijsbergen, C. J. 1999. *The Impact on Retrieval Effectiveness of Skewed Frequency Distributions,* ACM Transactions in Information Systems 17(4), ACM Press, Pp. 440 – 465.

Sanderson, M. 2000. *Retrieving with Good Sense* in Information Retrieval 2(1), Kluwer Academic Publishing, Pp. 49 – 69.

Schütze, H. 1998. *Automatic Word Sense Disambiguation,* Computational Linguistics 24(1), Pp. 113-120.

Schütze, H and Pederson, J. O. 1995. *Information Retrieval Based on Word Senses*, in Proceedings of the 4th Symposium on Document Analysis and Information Retrieval, Pp. 161 -175.

Stokoe, C. M; Oakes, M J; Tait, J I. 2003. *Word Sense Disambiguation in Information Retrieval Revisited* in Proceedings of the 26th ACM SIGIR, Toronto, Canada, Pp. 159 – 166.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least-Effort*, Cambridge MA, U.S.A., Addison-Wesley