Lexical ambiguity and Information Retrieval revisited

Julio Gonzalo Anselmo Peñas Felisa Verdejo

UNED
Ciudad Universitaria, s.n.
28040 Madrid - Spain
{julio,anselmo,felisa}@ieec.uned.es

Abstract

A number of previous experiments on the role of lexical ambiguity in Information Retrieval are reproduced on the IR-Semcor test collection (derived from Semcor), where both queries and documents are hand-tagged with phrases, Part-Of-Speech and WordNet 1.5 senses.
Our results indicate that a) Word Sense Disambiguation can be more beneficial to Information Retrieval than the experiments of Sanderson (1994) with artificially ambiguous pseudo-words suggested. b) Part-Of-Speech tagging does not seem to help improving retrieval, even if it is manually annotated. c) Using phrases as indexing terms is not a good strategy if no partial credit is given to the phrase components.

1 Introduction

A major difficulty to experiment with lexical ambiguity issues in Information Retrieval is always to differentiate the effects of the indexing and retrieval strategy being tested from the effects of tagging errors. Some examples are:

1. In (Richardson and Smeaton, 1995), a sophisticated retrieval system based on conceptual similarity resulted in a decrease of IR performance. It was not possible, however, to distinguish the effects of the strategy and the effects of automatic Word Sense Disambiguation (WSD) errors. In (Smeaton and Quigley, 1996), a similar strategy and a combination of manual disambiguation and very short documents -image captions- produced, however, an improvement of IR performance.
2. In (Krovetz, 1997), discriminating word senses with different Part-Of-Speech (as annotated by the Church POS tagger) also harmed retrieval efficiency. Krovetz noted than more than half of the words in a dictionary that differ in POS are related in meaning, but he could not decide whether the decrease of performance was due to the loss of such semantic relatedness or to automatic POS tagging errors.
3. In (Sanderson, 1994), the problem of discerning the effects of differentiating word senses from the effects of inaccurate disambiguation was overcome using artificially created pseudo-words (substituting, for instance, all occurrences of *banana* or *kalashnikov* for *banana/kalashnikov*) that could be disambiguated with 100% accuracy (substituting *banana/kalashnikov* back to the original term in each occurrence, either *banana* or *kalashnikov*). He found that IR processes were quite resistant to increasing degrees of lexical ambiguity, and that disambiguation harmed IR efficiency if performed with less that 90% accuracy. The question is whether real ambiguous words would behave as pseudo-words.
4. In (Schütze and Pedersen, 1995) it was shown that sense discriminations extracted from the test collections may enhance text retrieval. However, the static sense inventories in dictionaries or thesauri -such as WordNet- have not been used satisfactorily in IR. For instance, in (Voorhees, 1994), manual expansion of TREC queries with semantically related words from WordNet only produced slight improvements with the shortest queries.

In order to deal with these problems, we designed an IR test collection which is hand annotated with Part-Of-Speech and semantic tags from WordNet 1.5. This collection was first introduced in (Gonzalo et al., 1998) and it is described in Section 2. This collection is quite small for current IR standards (it is only slightly bigger than the TIME collection), but offers a unique chance to analyze the behavior of semantic approaches to IR before scaling them up to TREC-size collections (where manual tagging is unfeasible).

In (Gonzalo et al., 1998), we used the manual annotations in the IR-Semcor collection to show that indexing with WordNet synsets can give significant improvements to Text Retrieval, even for large queries. Such strategy works better than the synonymy expansion in (Voorhees, 1994), probably because it identifies synonym terms but, at the same

time, it differentiates word senses.

In this paper we use a variant of the IR-Semcor collection to revise the results of the experiments by Sanderson (Sanderson, 1994) and Krovetz (Krovetz, 1997) cited above. The first one is reproduced using both ambiguous pseudo-words and real ambiguous words, and the qualitative results compared. This permits us to know if our results are compatible with Sanderson experiments or not. The effect of lexical ambiguity on IR processes is discussed in Section 3, and the sensitivity of recall/precision to Word Sense Disambiguation errors in Section 4. Then, the experiment by Krovetz is reproduced with automatic and manually produced POS annotations in Section 5, in order to discern the effect of annotating POS from the effect of erroneous annotations. Finally, the richness of multiwords in WordNet 1.5 and of phrase annotations in the IR-Semcor collection are exploited in Section 6 to test whether phrases are good indexing terms or not.

2 The IR-SEMCOR test collection

The best-known publicly available corpus hand-tagged with WordNet senses is SEMCOR (Miller et al., 1993), a subset of the Brown Corpus of about 100 documents that occupies about 2.4 Mb. of text (22Mb. including annotations). The collection is rather heterogeneous, covering politics, sports, music, cinema, philosophy, excerpts from fiction novels, scientific texts...

We adapted SEMCOR in order to build a test collection -that we call IR-SEMCOR- in four manual steps:

- We have split the documents in Semcor 1.5 to get coherent chunks of text for retrieval. We have obtained 171 fragments with an average length of 1331 words per fragment. The new documents in Semcor 1.6 have been added without modification (apart from mapping Wordnet 1.6 to WordNet 1.5 senses), up to a total of 254 documents.
- We have extended the original *TOPIC* tags of the Brown Corpus with a hierarchy of sub-tags, assigning a set of tags to each text in our collection. This is not used in the experiments reported here.
- We have written a summary for each of the first 171 fragments, with lengths varying between 4 and 50 words and an average of 22 words per summary. Each summary is a human explanation of the text contents, not a mere bag of related keywords.
- Finally, we have hand-tagged each of the summaries with WordNet 1.5 senses. When a word or term was not present in the database, it was left unchanged. In general, such terms correspond to proper nouns; in particular, groups (vg. *Fulton_County_Grand_Jury*), persons (*Cervantes*) or locations (*Fulton*).

We also generated a list of "stop-senses" and a list of "stop-synsets", automatically translating a standard list of stop words for English.

In our first experiments (Gonzalo et al., 1998; Gonzalo et al., 1999), the summaries were used as queries, and every query was expected to retrieve exactly one document (the one summarized by the query). In order to have a more standard set of relevance judgments, we have used the following assumption here: if an original Semcor document was split into n chunks in our test collection, the summary of each of the chunks should retrieve all the chunks of the original document. This gave us 82 queries with an average of 6.8 relevant documents per query. In order to test the plausibility of this artificial set of relevance judgments, we produced an alternative set of random relevance judgments. This is used as a baseline and included for comparison in all the results presented in this paper.

The retrieval engine used in the experiments reported here is the INQUERY system (Callan et al., 1992).

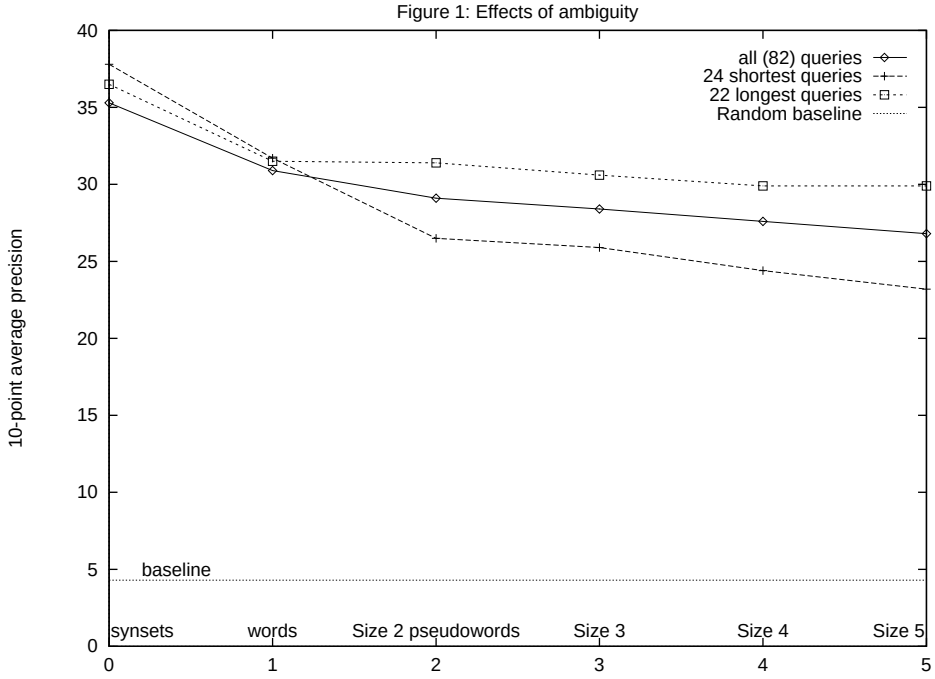
3 Lexical Ambiguity and IR

Sanderson used a technique previously introduced in (Yarowski, 1993) to evaluate Word Sense Disambiguators. Given a text collection, a (size 2) pseudo-word collection is obtained by substituting all occurrences of two randomly chosen words (say, *bank* and *spring*) by a new ambiguous word (*bank/spring*). Disambiguating each occurrence of this pseudo-word consists on finding whether the original term was either *bank* or *spring*. Note that we are not strictly discriminating senses, but also conflating synonym senses of different words. We previously showed (Gonzalo et al., 1998) that WordNet synsets seem better indexing terms than senses.

Sanderson used an adapted version of the Reuters text categorization collection for his experiment, and produced versions with pseudo-words of size 2 to 10 words per pseudo-word. Then he evaluated the decrease of IR performance as the ambiguity of the indexing terms is increased. He found that the results were quite insensitive to ambiguity, except for the shortest queries.

We have reproduce Sanderson's experiment for pseudo-words ranging from size 1 (unmodified) to size 5. But when the pseudo-word *bank/spring* is disambiguated as *spring*, this term remains ambiguous: it can be used as *springtime*, or *hook*, or *to jump*, etc. We have, therefore, produced another collection of "ambiguity 0", substituting each word by its WordNet 1.5 semantic tag. For instance, *spring* could be

Figure 1: Effects of ambiguity



substituted for $n07062238$, which is a unique identifier for the synset {*spring, springtime: the season of growth*}.

The results of the experiment can be seen in Figure 1. We provide 10-point average precision measures [1] for ambiguity 0 (synsets), 1 (words), and 2 to 5 (pseudo-words of size 2,3,4,5). Three curves are plotted: all queries, shortest queries, and longer queries. It can be seen that:

- The decrease of IR performance from synset indexing to word indexing (the slope of the leftmost part of the figure) is more accused than the effects of adding pseudoword ambiguity (the rest of the figure). Thus, reducing real ambiguity seems more useful than reducing pseudo-word ambiguity.
- The curve for shorter queries have a higher slope, confirming that resolving ambiguity is more benefitial when the relative contribution of each query term is higher. This is true both for real ambiguity and pseudo-word ambiguity. Note, however, that the role of real ambiguity is more important for longer queries than pseudo-word ambiguity: the curve for longer queries has a high slope from synsets to words, but it is very smooth from size 1 to size 5 pseudo-words.
- In our experiments, shorter queries behave better than longer queries for synset indexing (the leftmost points of the curves). This unexpected behavior is idiosyncratic of the collection: our documents are fragments from original Semcor texts, and we hypothesize that fragments of one text are relevant to each other. The shorter summaries are correlated with text chunks that have more cohesion (for instance, a Semcor text is split into several IRSemcor documents that comment on different baseball matches). Longer summaries behave the other way round: IRSemcor documents correspond to less cohesive text chunks. As introducing ambiguity is more harming for shorter queries, this effect is quickly shadowed by the effects of ambiguity.

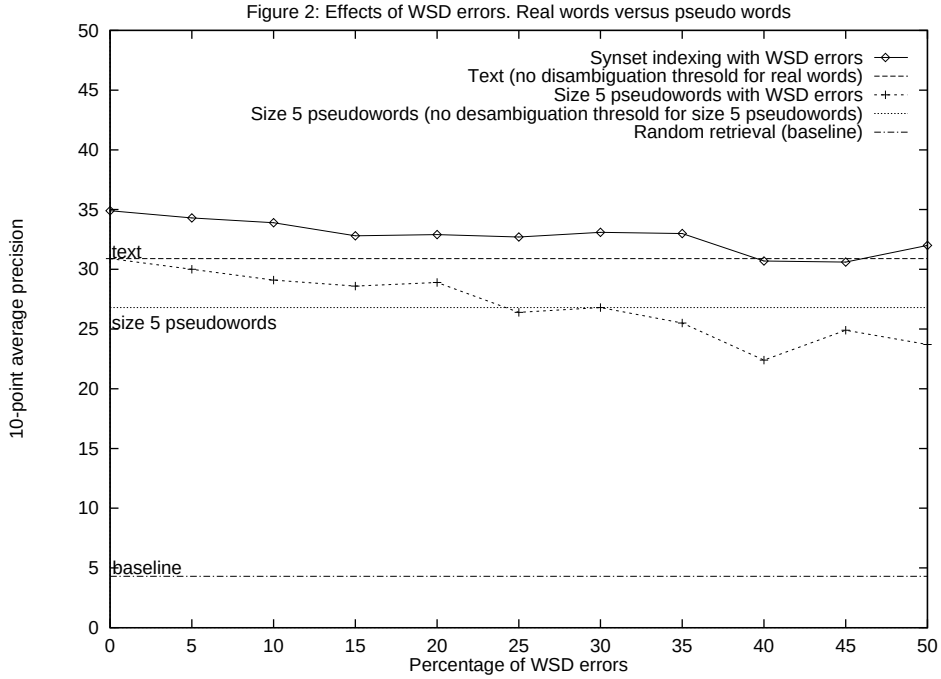
[1] The 10-point average precision is a standard IR measure obtained by averaging precision at recall points $10, 20, \ldots 100$.

4 WSD and IR

The second experiment carried out by Sanderson was to disambiguate the size 5 collection introducing fixed error rates (thus, the original pseudo-word collection would correspond to 100% correct disambiguation). In his collection, disambiguating below 90% accuracy produced worse results than not disambiguating at all. He concluded that WSD needs to be extremely accurate to improve retrieval results rather than decreasing them.

We have reproduce his experiment with our size 5 pseudo-words collection, ranging from 0% to 50% error rates (100% to 50% accuracy). In this case, we have done a parallel experiment performing real Word Sense Disambiguation on the original text collection, introducing the fixed error rates with respect to the manual semantic tags. The error rate is understood as the percentage of polysemous words in-

Figure 2: Effects of WSD errors. Real words versus pseudo words



correctly disambiguated.

The results of both experiments can be seen in Figure 2. We have plotted 10-point average precision in the Y-axis against increasing percentage of errors in the X-axis. The curve representing real WSD has as a threshold the 10-pt average precision for plain text, and the curve representing pseudo-disambiguation on the size-5 pseudo-word collection has as threshold the results for the size-5 collection without disambiguation. From the figure it can be seen that:

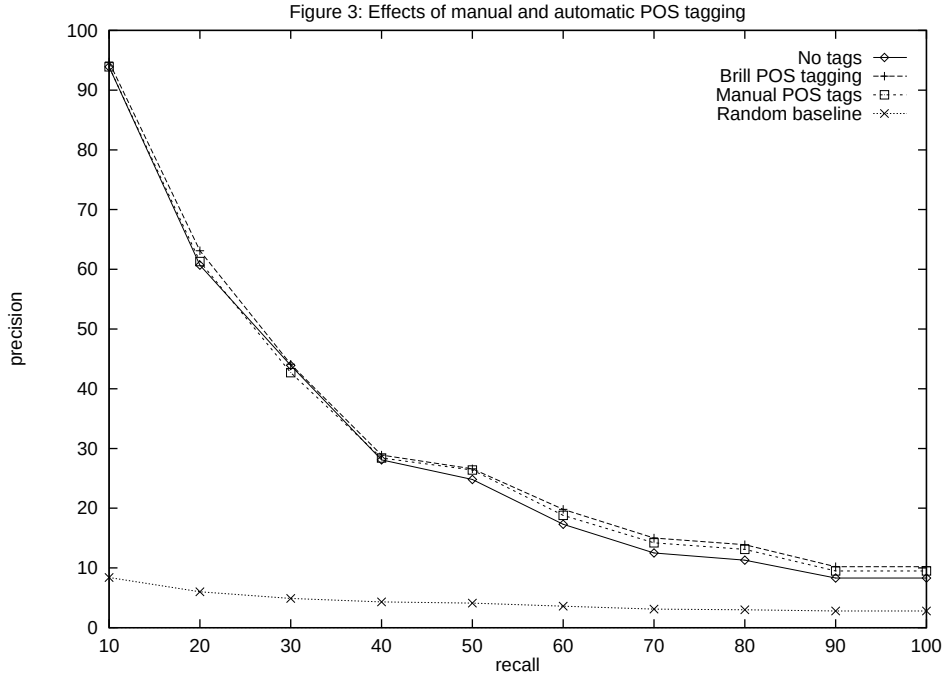
- In the experiment with size 5 pseudo-word disambiguation, our collections seems to be more resistant to WSD errors than the Reuters collection. The 90% accuracy threshold is now 75%.
- The experiment with real disambiguation is more tolerant to WSD errors. Above 60% accuracy (40% error rate) it is possible to improve the results of retrieval with plain text.

The discrepancy between the behavior of pseudo-words and real ambiguous terms may reside in the nature of real polysemy:

- Unlike the components of a pseudo-word, the different meanings of a real, polysemous word are often related. In (Buitelaar, 1998) it is estimated that only 5% of the word stems in WordNet can be viewed as true homonyms (unrelated senses), while the remaining 95% polysemy can be seen as predictable extensions of a core sense (*regular polysemy*). Therefore, a disambiguation error might be less harmful if a strongly related term is chosen. This fact also suggests that Information Retrieval does not necessarily demand full disambiguation. Rather than picking one sense and discarding the rest, WSD in IR should probably weight senses according to their plausibility, discarding only the less likely ones. This is used in (Schütze and Pedersen, 1995) to get a 14% improvement of the retrieval performance disambiguating with a co-occurrence-based induced thesaurus. This is an issue that arises naturally when translating queries for Cross-Language Text Retrieval, in contrast to Machine Translation. A Machine Translation system has to choose one single translation for every term in the source document. However, a translation of a query in Cross-Language retrieval has to pick up all likely translations for each word in the query. In (Gonzalo et al., 1999) we argue that mapping a word into word senses (or WordNet synsets) is strongly related to that problem.

- Although the average polysemy of the terms in the Semcor collection is around five (as in Sanderson's experiment), the average polysemy of WordNet 1.5 terms is between 2 and 3. The reason is that polysemy is correlated with frequency of usage. That means that the best discriminators for a query will be (in general) the less polysemous terms. The more polysemous terms are more frequent and thus worse discriminators, and disambiguation errors are not

Figure 3: Effects of manual and automatic POS tagging



as harmful as for the pseudo-words experiment.

5 POS tagging and IR

Among many other issues, Krovetz tested to what extent Part-Of-Speech information was a good source of evidence for sense discrimination. He annotated words in the TIME collection with the Church Part-Of-Speech tagger, and found that performance decreased. Krovetz was unable to determine whether the results were due to the tagging strategy or to the errors made by the tagger. He observed that, in many cases, words were related in meaning despite a difference in Part-Of-Speech (for instance, in "summer shoes *design*" versus "they *design* sandals"). But he also found that not all errors made by the tagger cause a decrease in retrieval performance.

We have reproduced the experiment by Krovetz in our test collection, using the Brill POS tagger, on one hand, and the manual POS annotations, on the other. The precision/recall curves are plotted in Figure 3 against plain text retrieval. That curves does not show any significant difference between the three approaches. A more detailed examination of some representative queries is more informative:

5.1 Manual POS tagging vs. plain text

Annotating Part-Of-Speech misses relevant information for some queries. For instance, a query containing "*talented baseball player*" can be matched against a relevant document containing "*is one of the top talents of the time*", because stemming conflates *talented* and *talent*. However, POS tagging gives *ADJ/talent* versus *N/talent*, which do not match. Another example is "*skilled diplomat of an Asian Country*" versus "*diplomatic policy*", where *N/diplomat* and *ADJ/diplomat* are not matched.

However, the documents where the matching terms agree in category are ranked much higher with POS tagging, because there are less competing documents. The two effects seem to compensate, producing a similar recall/precision curve on overall.

Therefore, annotating Part-Of-Speech does not seem worthy as a standalone indexing strategy, even if tagging is performed manually. Perhaps giving partial credit to word occurrences with different POS would be an interesting alternative.

Annotating POS, however, can be a useful intermediate task for IR. It is, for instance, a first step towards semantic annotation, which produced much better results in our experiments.

5.2 Brill vs. manual tagging

Although the Brill tagger makes more mistakes than the manual annotations (which are not error free anyway), the mistakes are not completely correlated to retrieval decrease. For instance, a query about "*summer shoe design*" is manually annotated as "*summer/N shoe/N design/N*", while the Brill tagger produces "*summer/N shoe/N design/V*". But an appropriate document contains "*Italian designed sandals*", which is manually annotated as "*Italian/ADJ designed/ADJ sandals/N*" (no match), but as "*Italian/ADJ designed/V sandals/N*" by the Brill tagger (matches design and designed after stemming).

In general, comparing with no tagging, the automatic and the manual tagging behave in a very similar way.

6 Phrase indexing

WordNet is rich in multiword entries (more than 55000 variants in WordNet 1.5). Therefore, such collocations are annotated as single entries in the Semcor and IR-Semcor collections. The manual annotation also includes name expressions for persons, groups, locations, institutions, etc., such as *Drew Centennial Church* or *Mayor-nominate Ivan Allen Jr.*. In (Krovetz, 1997), it is shown that the detection of phrases can be useful for retrieval, although it is crucial to assign partial credit also to the components of the collocation.

We have performed an experiment to compare three different indexing strategies:

1. Use plain text both for documents and queries, without using phrase information.
2. Use manually annotated phrases as single indexing units in documents and queries. This means that *New_York* is a term unrelated to *new* or *York* (which seems clearly beneficial both for weighting and retrieval), but also that *Drew_Centennial_Church* would be a single indexing term unrelated to *church*, which can lead to precise matchings, but also to lose correct query/document correlations.
3. Use plain text for documents, but exploit the INQUERY `#phrase` query operator for the collocations in the query. For instance, *meeting of the National_Football_League* is expressed as `#sum(meeting #phrase(National Football League))` in the query language. The `#phrase` operator assigns credit to the partial components of the phrase, while priming its co-occurrence.

The results of the experiments can be seen in Figure 4. Overall, indexing with multiwords behaves slightly worse than standard word indexing. Using the INQUERY `#phrase` operator behaves similarly to word indexing.

A closer look at some case studies, however, gives more information:

- In some cases, simply indexing with phrases is obviously the wrong choice. For instance, a query containing "*candidate in governor's_race*" does not match "*opened his race for governor*". This supports the idea that it is crucial to assign credit to the partial components of a phrase, and also that it may be useful to look for co-occurrence beyond one word windows.
- Phrase indexing works much better when the query is longer and there are relevant terms apart from one or more multiwords. In such cases, a relevant document containing just one query term is ranked much higher with phrase indexing, because false partial matches with a phrase are not considered. Just using the `#phrase` operator behaves mostly like no phrase indexing for these queries, because this filtering is not achieved.
- Phrase indexing seems more adequate when the query is intended to be precise, which is not the case of our collection (we assume that the summary of a fragment has all the fragments in the original text as relevant documents). For instance, "*story of a famous strip cartoonist*" is not related -with phrase indexing- to a document containing "*detective_story*". This is correct if the query is intended to be strict, although in our collection these are fragments of the same text and thus we are assuming they are related. The same happens with the query "*The board_of_regents of Paris_Junior_College has named the school's new president*", which is not related to "*Junior or Senior High School Teaching Certificate*". This could be the right decision in a different relevance judgment setup, but it is wrong for our test collection.

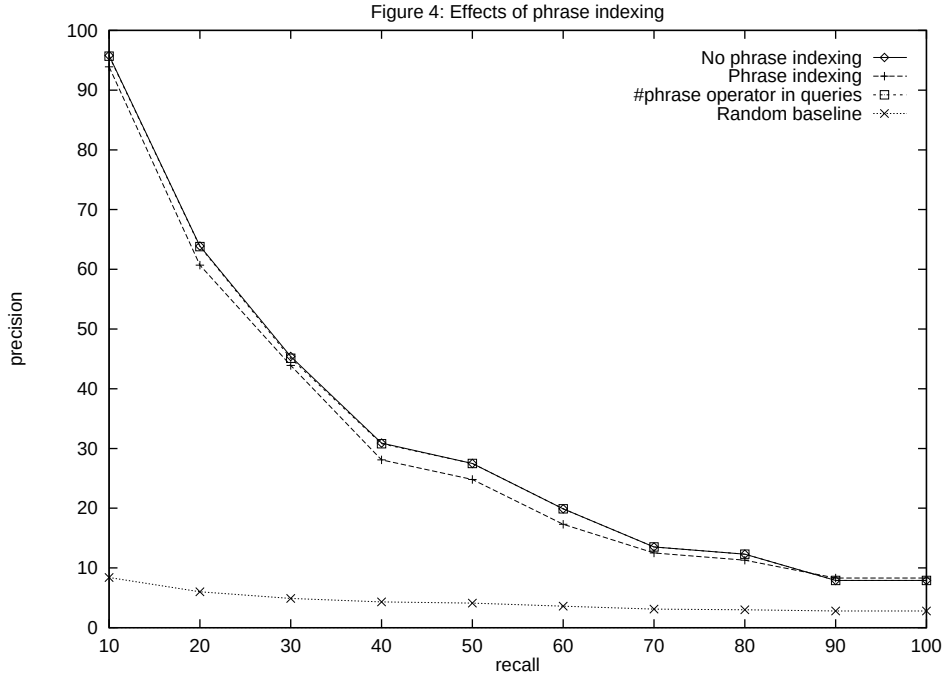
7 Conclusions

We have revised a number of previous experiments regarding lexical ambiguity and Information Retrieval, taking advantage of the manual annotations in our IR-Semcor collection. Within the limitations of our collection (mainly its reduced size), we can extract some conclusions:

- Sense ambiguity could be more relevant to Information Retrieval than suggested by Sanderson's experiments with pseudo-words. In particular, his estimation that 90% accuracy is needed to benefit from Word Sense Disambiguation techniques does not hold for real ambiguous words in our collection.
- Part-Of-Speech information, even if manually annotated, seems too discriminatory for Information Retrieval purposes. This clarifies the results obtained by Krovetz with an automatic POS tagger.
- Taking phrases as indexing terms may decrease retrieval efficiency. Phrase indexing could be more useful, anyway, when the queries demands a very precise kind of documents, and when the number of available documents is high.

In our opinion, lexical ambiguity will become a central topic for Information Retrieval as the importance of Cross-Language Retrieval grows (something

Figure 4: Effects of phrase indexing



that the increasing multilinguality of Internet is already producing). Although the problem of Word Sense Disambiguation is still far from being solved, we believe that specific disambiguation for (Cross-Language) Information Retrieval could achieve good results by weighting candidate senses without a special commitment to Part-Of-Speech differentiation. An interesting point is that the WordNet structure is not well suited for IR in this respect, as it keeps noun, verb and adjective synsets completely unrelated. The EuroWordNet multilingual database (Vossen, 1998), on the other hand, features cross-part-of-speech semantic relations that could be useful in an IR setting.

Acknowledgments

Thanks to Douglas Oard for the suggestion that originated this work.

References

P. Buitelaar. 1998. *CoreLex: systematic polysemy and underspecification.* Ph.D. thesis, Department of Computer Science, Brandeis University, Boston.

J. Callan, B. Croft, and S. Harding. 1992. The INQUERY retrieval system. In *Proceedings of the 3rd Int. Conference on Database and Expert Systems applications.*

J. Gonzalo, M. F. Verdejo, I. Chugur, and J. Cigarrán. 1.998. Indexing with Wordnet synsets can improve Text Retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems.*

J. Gonzalo, F. Verdejo, and I. Chugur. 1999. Using EuroWordNet in a concept-based approach to Cross-Language Text Retrieval. *Applied Artificial Intelligence, Special Issue on Multilinguality in the Software Industry: the AI contribution.*

R. Krovetz. 1997. Homonymy and polysemy in Information Retrieval. In *Proceedings of ACL/EACL'97.*

G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology.* Morgan Kauffman.

R. Richardson and A.F. Smeaton. 1995. Using Wordnet in a knowledge-based approach to Information Retrieval. In *Proceedings of the BCS-IRSG Colloquium, Crewe.*

M. Sanderson. 1994. Word Sense Disambiguation and Information Retrieval. In *Proceedings of* 17^th *International Conference on Research and Development in Information Retrieval.*

H. Schütze and J. Pedersen. 1995. Information Retrieval based on word senses. In *Fourth Annual Symposium on Document Analysis and Information Retrieval.*

A.F. Smeaton and A. Quigley. 1996. Experiments on using semantic distances between words in image caption retrieval. In *Proceedings of the* 19^{th} *International Conference on Research and Development in Information Retrieval.*

Ellen M. Voorhees. 1994. Query expansion using

lexical-semantic relations. In *Proceedings of the* 17^th *International Conference on Research and Development in Information Retrieval.*

Vossen, P. (ed). 1998. *EuroWordNet: a multilingual database with lexical semantic networks.* Kluwer Academic Publishers.

D. Yarowski. 1993. One sense per collocation. In *Proceedings of ARPA Human Language Technology Workshop.*