

An Experimental Study on Syntactic and Semantic Annotations in Text Retrieval

George Tsatsaronis
Biotechnology Center, Technische Universität Dresden, Dresden, Germany
george.tsatsaronis@biotec.tu-dresden.de

ABSTRACT

Syntactic and semantic ambiguity affect the text retrieval task as each type of ambiguity influences precision and/or recall. In this work we provide an experimental study on the effect of several ambiguity types in IR, by resolving each ambiguity type separately and adding respective annotations in the indexed text. We focus on five small text retrieval collections.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Retrieval Models*

General Terms

Algorithms, Experimentation.

1. INTRODUCTION

Although lexical ambiguity is recognized as a problem in IR, little work has been done to estimate the effect of each ambiguity type independently and the role of respective annotations in retrieval performance. We experimentally evaluate the effect of three different annotation types to text retrieval, namely: *part of speech tags*, *word senses*, and *identification of phrases*. Among the related studies made in the past, the works by Krovetz Voorhees [4], Sanderson and van Rijsbergen [2], and Stokoe [3] are representative. Voorhees used a real word sense disambiguation (WSD) system and the rest a methodology based on the generation of random pseudowords to assess the role of semantic ambiguity in IR. The current study is a complementary work to the aforementioned; it investigates the relation between each ambiguity type and the retrieval performance independently and analyzes the effect that the respective annotations may have to the retrieval performance.

With regards to the types of lexical ambiguity, they can be categorized in *syntactic* and *semantic*. *Syntactic* ambiguity occurs from the different syntactic roles that a word may have in context and can be resolved with *POS tagging*. Orthogonal to *syntactic*, is *semantic* ambiguity, which occurs from *polysemy* or *homonymy* of words. In this study we consider both as a single type of semantic ambiguity, namely *sense ambiguity*, which can be resolved with WSD. Furthermore, a word can occur as part of a phrase, e.g., the words *square* and *root* can occur individually, but also as part of the phrase *square root* in *WordNet*. This can be considered as a type of semantic ambiguity and can be resolved with *phrase recognition*, e.g., implementing a dictionary look-up. Finally, *stemming* may be

considered as the process of mapping a lemma into its most general lexeme, and can, thus, be regarded as a resolution to semantic ambiguity.

2. LEXICAL ANNOTATIONS

In all experiments to follow the *vector space model* (VSM) is used for documents and queries representation. The conventional $TF \cdot IDF$ is used as a terms' weighting scheme and *cosine* as a similarity between vectors. For every type of ambiguity we altered the VSM representation of both documents and queries, to incorporate annotations produced by the resolution of the respective ambiguity type. For the *syntactic* ambiguity, we considered a space where each term is indexed along with its *POS* tag, e.g., each term occurrence t_i is indexed as $t_i_POS_i$, where POS_i is the respective *POS* tag. *POS* tags are found using the *Stanford Log-Linear POS Tagger*. For *sense ambiguity*, we created 3 different *Generalized Vector Space Models* (GVSM), which embed senses information. All queries were manually disambiguated and a baseline WSD system is applied to the documents, selecting always the first sense from *WordNet*. The found senses are added in the vectors, along with the terms. The difference in the 3 used GVSM lies in the weighting of terms and senses. *GVSM1* considers two different vector spaces; terms and senses. Each one uses its own, separate $TF \cdot IDF$ weighting, and the final similarity between a document and a query is computed as the sum of the cosine similarities in the two spaces. *GVSM2* considers terms and senses in the same space. *GVSM3* [1], considers one hybrid vector space of terms and senses where senses are assigned with the $TF \cdot IDF$ weight of their terms. To study the effect of *phrase detection* we perform a dictionary look-up in *WordNet* using a sliding window of varying length, starting from 7 terms and dropping down to 2. The resulting phrases substitute the respective term occurrences in the indexing. Finally, the effect of stemming is studied by applying the *Porter stemmer*. The stems substitute the respective term occurrences in the indexing.

3. EXPERIMENTAL EVALUATION

3.1 Description of the Data Collections

The evaluation was conducted in 5 known IR collections shown in Table 1, which reports: the collection domain, number of documents (D.) and queries (Q.), average number of term occurrences in documents (#T D.) and queries (#T Q.), average number of phrases recognized in documents (#P D.) and queries (#P Q.), percentages of documents and queries in which at least one phrase was recognized from the dictionary, and, the average ambiguity of words found in *WordNet* for all documents (D. amb.) and queries (Q.

	D.	Q.	Domain	#T D.	#T Q.	#P D.	#P Q.	D. amb.	Q. amb.				
CACM	3204	64	ACM abstr.	29.5	13.4	2.11	49.7%	1.4	39%	5	88.3%	4.5	84.9%
MED.	1033	30	med. abstr.	82.3	11.6	2.9	89%	1.3	66.6%	4.4	82.2%	3.1	90.3%
TIMES	423	83	general	304.6	8.2	9.8	100%	1.3	57.8%	4.5	70.5%	3.6	77.9%
NPL	11429	100	physics	23.4	6.8	1.4	44.9%	1.2	26%	4.6	70.2%	4.3	62.1%
CRAN.	1400	225	aer/mics	90.4	9.3	3.1	88%	1.1	35.1%	5.3	88.1%	4.7	89.9%

Table 1: Documents, queries and domains of the retrieval collections.

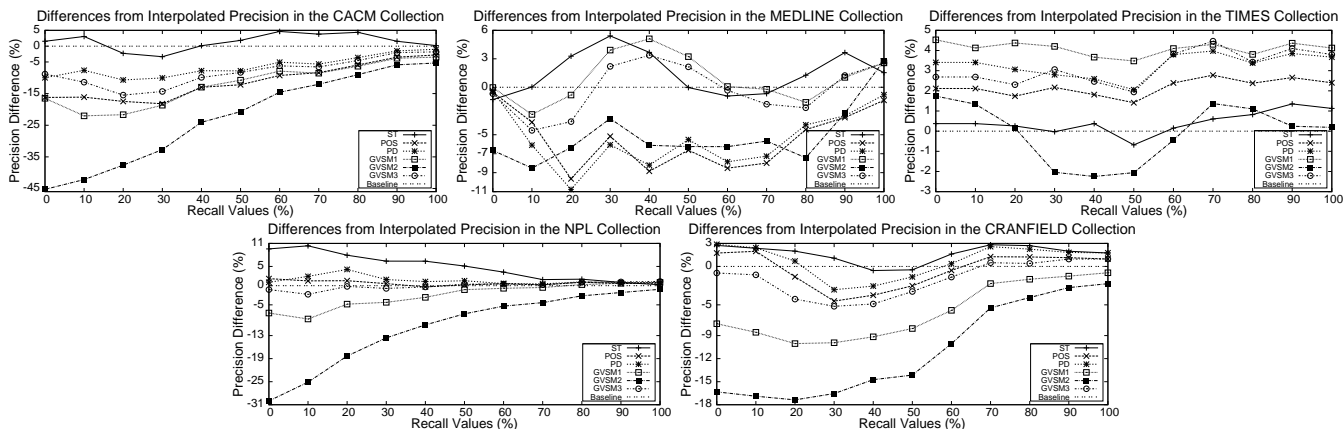


Figure 1: Differences from the baseline in interpolated precision.

amb.), along with the percentage of terms found in the used lexicon.

3.2 Results and Analysis

Figure 1 presents the differences of each retrieval model from the interpolated precision of the *VSM* baseline (indexed terms without resolving any ambiguity) for the 11-standard recall points. Each line is one of the six retrieval models, namely *stemming* (*ST*), *POS tagging* (*POS*), *phrase detection* (*PD*), *GVSMI*, *GVSMD* and *GVSMD3*.

Stems: As shown, stemming improves almost in all cases the IR performance. Initially it boosts precision for the first 2 or 3 recall points, then it weakens in the 40% – 60% recall levels, and finally it boosts precision again, until all relevant documents are retrieved. Looking closely at *NPL*, we see that it boosted precision constantly (up to 11% p.p.), in all recall points. Since *NPL* has the smallest documents, we may infer that stemming may boost precision more in cases when documents are relatively small.

POS Tags: *POS Tags* cannot improve much precision. The respective model can add as much as 2.5% p.p., and this only happens in *TIMES*. In *CRANFIELD* it boosts precision by 2% p.p., but in the same collection it drops up to 5% p.p. in the medium recall points. In *TIMES* and *CACM*, in which the model achieved its top and its worst performance respectively, we notice that *TIMES* has the largest documents and *CACM* is among the two collections with the shortest. Thus, *POS* tags cannot boost precision, but in the few cases it des, the collections have relatively large documents.

Phrases: *Phrase detection* can boost precision up to 4% p.p. in the tested collections. In all recall levels, this happens only in *TIMES* and *NPL*. In *CACM*, the effect is negative, as it drops precision by almost 10% p.p. in the first 3 recall points. From Table 1, we see that *TIMES* has many phrases detected on average per document (9.8), while *CACM* very few (2.1). Thus, *phrase detection* requires the occurrence of many phrases in documents to aid IR performance.

Sense tags: The *GVSMDs* used show that embedding *WSD* information in retrieval can boost precision up to 5% p.p., but can also drop it by more than 45%. *GVSMD1* and *GVSMD3* have similar behavior, and are both from *GVSMD2*. In general, considering a single vector space and mixing up term and sense dimensions seems a bad choice. The two collections that sense tags may improve precision, *TIMES* and *MEDLINE*, are the collections with the less ambiguous terms, both per document and per query. Thus, in cases where the average ambiguity of the disambiguated terms is relatively low, sense tags may help. However, further study is needed to investigate how the *WSD* performance affects these results.

4. CONCLUSIONS

In this work, we studied the effect of several tag types in IR performance. the tags are produced by resolving four types of syntactic and semantic ambiguity. Our results indicate that *stems*, and *sense* tags can boost precision at almost every recall level, under conditions, while *POS* tags cannot. In the future we will investigate the effect of combining these tags to improve IR.

5. REFERENCES

- [1] D. Mavroudis, G. Tsatsaronis, M. Vazirgiannis, M. Theobald, and G. Weikum. Word sense disambiguation for exploiting hierarchical thesauri in text classification. In *Proc. of PKDD*, pages 181–192, 2005.
- [2] M. Sanderson and C. van Rijsbergen. The impact on retrieval effectiveness of skewed frequency distributions. *ACM TIS*, 17(4):440–465, 1999.
- [3] C. Stokoe. Differentiating homonymy and polysemy in information retrieval. In *Proc. of HLT/EMNLP*, pages 403–410, 2005.
- [4] E. Voorhees. Using wordnet to disambiguate word sense for text retrieval. In *Proc. of SIGIR*, pages 171–180, 1993.