Learning to Augment a Machine-Readable Dictionary

Robert Krovetz
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
USA

Abstract

Dictionaries will always be incomplete; sometimes a word will acquire a new sense in a technical field, and new words are being added to the language all the time. This paper will discuss our comparisons between a machine-readable dictionary and various information retrieval test collections. We will first report on the number of words found in the dictionary, and how much improvement is gained by going to a larger dictionary. We will then discuss experiments concerned with augmenting the dictionary with information acquired from the corpus, and by exploiting redundancy within the dictionary itself.

1 Introduction

Dictionaries will always be incomplete; sometimes a word will acquire a new sense in a technical field, and new words are being added to the language all the time. While it is clear that dictionaries need to be supplemented with information from corpora, relatively little quantitative information is available about the extent of the gap. How good is the dictionary's coverage of the language? How much improvement is gained by going from a small dictionary to a large one? To answer these questions we examined the lexicons of four different test collections used in information retrieval. We determined how many words were found in the *Longman Dictionary of Contemporary English* [Proctor 78], and how many of the words not found would appear in a larger dictionary, the *Collins English Dictionary*. We also conducted experiments to determine gaps in the dictionary with respect to part-of-speech, morphology, and subject-area codes (these are codes that are associated with some

	CACM	TIME	NPL	WEST
Number of queries	64	83	93	34
Number of documents	3204	423	11429	11953
Mean words per query	13.0	8.9	7.1	9.6
Mean words per document	62	581	43	3262
Mean relevant documents per query	15.84	3.90	22.3	28.9
Number of words in collection	200,000	250,000	490,000	39,000,000

Table 1: Statistics on information retrieval test collections. Each collection represents a different subject area. CACM is about Computer Science, TIME is primarily about politics (*Time Magazine*), NPL is about physics, and WEST is about law.

of the senses in the machine-readable version of *Longman*; they will be described in more detail later in the paper). Our aim was to get a better understanding of the coverage of a machine-readable dictionary, and the extent to which gaps in the lexicon could be augmented with information from the corpora. In addition, differences between corpora and dictionaries can be associated with differences in word meaning (e.g., *reciprocal* as an adjective or as a noun). We wanted to determine how often this was the case, and what problems would be encountered in an effort to automatically update the dictionary with new word meanings. The following section will provide statistics about the corpora used in our experiments, and we will then describe the experiments themselves.

2 Collection Statistics

The test collections are text databases that are used as a standard for assessing performance in the information retrieval field. They consist of a set of documents, a set of queries, and relevance judgements that indicate which documents are relevant to each query. Each collection covers a different domain (computer science, newspaper stories, physics, and law), and they represent a wide range in terms of average document length and overall number of documents. The statistics for the different test collections are given in Table 1.

	CACM	TIME	NPL	WEST
Numbers	5.5/1.9	4.0/3.6	0/0	17.8/11.6
Slashonyms	0/0	0/0	0/0	0.1/0.0
Contractions	0/0	0/0	0/0	0.6/0.2
Initialisms	0/0	0.0/0.7	0/0	1.5/2.4
Hyphenated	12.8/2.7	6.6/2.1	0/0	13.5/0.7
Proper Nouns	6.2/3.1	11.2/12.8	7.8/2.6	24.1/5.5
Longman	34.3/59.5	39.1/57.0	34.8/65.6	6.7/58.2
Short Words	2.0/1.0	0.8/1.1	3.9/1.0	1.1/1.3
Inflected	25.1/25.5	26.9/17.6	28.8/23.7	10.0/15.2
Derived	5.9/3.8	4.2/2.0	6.9/3.4	2.6/1.8
Collins	1.7/1.1	1.2/0.6	4.2/2.2	1.0/0.8
Capitalized	4.2/1.4	3.6/1.9	0/0	17.0/0.8
Unknown	2.3/0.0	2.4/0.6	13.5/1.5	4.0/1.5
Total	100/100	100/100	100/100	100/100

Table 2: Composition of the lexicon for information retrieval collections in terms of types/tokens. Each row indicates the percentage of the lexicon made up by the category after all the preceding categories have been removed.

3 Dictionary Coverage of Test Collections

The *Longman Dictionary* is a dictionary for learners of English as a second language. It contains approximately 27,000 non-phrasal headwords.[1] The *Collins English Dictionary* is a general purpose dictionary, and contains about 60,000 non-phrasal headwords.

The lexicon for each test collection was broken down into various categories: numbers, slashonyms (terms containing a slash), contractions, initialisms (terms containing embedded periods), hyphenated forms, proper nouns,[2] words in the *Longman Dictionary*, short words (3 letters or less which were not found in the dictionary; most of these are acronyms), inflectional variants, derivational variants, words in the *Collins English Dictionary* that were not in any of the previous categories, capitalized words that were not in any of the previous categories, and finally everything else. This breakdown was done in order to get a better understanding of the makeup of the various collections, and to see how the words in the different dictionaries fit into the overall lexicon. Table 2 lists the percentage of the lexicon which fell into each category, both in terms of unique words (types) as well as occurrences (tokens).

The statistics indicate that the words from *Longman* constitute about 35-40% of the types for the small collections, and about 60% of the tokens regardless of the collection size. Relatively little increase is seen by using a larger dictionary (*Collins* vs. *Longman*). We only gain an additional one percent (both in terms of types and tokens), with the exception of NPL. Most of the additional coverage comes from technical vocabulary (e.g., *dielectric*, *capacitor*, and *bandwidth* for NPL, *polynomial*, *recursion*, and *parameter* for CACM, and *supra*, *antitrust*, and *fiduciary*, for WEST). For TIME the primary increase came from locations that were not mentioned in the proper noun list; this is because the *Longman* dictionary does not include definitions for proper nouns.

4 Augmenting the Dictionary

The above figures only give a very coarse estimate of the coverage of a dictionary. To get a better estimate, we examined some of the information associated with a lexical entry: part-of-speech, morphology, and subject codes. We will discuss each of these in the following sections.

4.1 Part of Speech

To acquire information about part-of-speech gaps we tagged two of the test collections with a stochastic tagger[3] [Church 88], and then identified the words that were tagged with a part-of-speech that was not mentioned in the dictionary. We chose one technical collection (CACM) and one non-technical collection (TIME) to see if that made any difference. The aim of this experiment was to determine how often new (or related) word meanings could be identified by a difference in part-of-speech.

The CACM collection provided us with an initial list of 424 word/tag pairs.[4] Of these words, 106 were tagged as past-tense verbs, but *Longman* listed almost all of them as adjectives (the sole exception was *intended*, which was listed as a noun). An additional 104 words were tagged as present-tense verbs, but were listed in *Longman* as either nouns or adjectives. The Church tagger often fails to distinguish tensed verbs from adjectival participles and gerunds. This is also a task that is not easy for humans to accomplish, and tagged corpora have considerable variation in this area [Belmore 88]. The inconsistent tagging of participles/gerunds and tensed verbs would have resulted in a large number of false positives, so we eliminated these 210 pairs from further consideration.

The TIME collection yielded an initial list of 1143 word/tag pairs. Of these words, 546 were tagged as either past or present tense verbs, and were not analyzed further. A breakdown of the remaining differences for the two collections is given in Table 3.

	CACM		TIME	
Tagging error:	48	(22%)	176	(29%)
Participle:	40	(19%)	106	(18%)
Gerund:	10	(5%)	9	(2%)
Not a root:	29	(14%)	54	(9%)
Upper/Lower:	0	(0%)	34	(6%)
Longman error:	6	(3%)	20	(3%)
Unclear:	10	(5%)	16	(3%)
Misc error:	21	(10%)	56	(9%)
Zeromorph:	32	(15%)	116	(19%)
Domain sense:	18	(8%)	10	(2%)
Total:	214	(100%)	597	(100%)

Table 3: Differences between *Longman* part-of-speech and tagging

Most of the categories in Table 3 reflect various types of error, or cases that did not reflect a difference in meaning. The category *tagging error* means that the tag assigned by the tagger was incorrect. The *participle* and *gerund* categories indicate cases in which a word was tagged as an adjective or noun, but the root was listed in *Longman* as a verb. The *not a root* category means that the morphological analysis routines failed to find the correct root in the dictionary. The *Upper/Lower* category refers to errors caused by converting the case of the collection; originally the TIME collection was entirely in upper case, and the Church tagger would have tagged every word as a proper noun. The collection was converted to lower case, and any errors that were a result of that were recorded in this category. *Longman error* means that the dictionary did not have the correct part-of-speech; these were usually only found in the machine-readable version and had been corrected in the printed version. The category *Unclear* reflects differences in classification between *Longman* and the tagger in which it was difficult to determine which one was correct. Finally, *Miscellaneous errors* usually involved some bizarre context, or errors in the algorithm that was used, or cases that were hard to categorize.

The experiment was successful in identifying a number of cases of related or domain specific meanings. The category *zeromorph* refers to 'zero-affix' morphology, which means that the senses are related even though they differ in part-of-speech; in TIME they were typically noun/adjective ambiguities that fell into predictable classes (e.g., person/role relationships such as *deputy* and *volunteer*, or person/attribute relationships such as *brunette* and *giant*),

and in CACM they were either verbs that were being used as nouns (e.g., *transform*, *merge*, and *fetch*), or noun/adjective ambiguities similar to the ones that occurred in TIME. Domain specific meanings are indicated by the category *domain sense*. For CACM these were words like *shear* (an adjective used in computer graphics to describe an angle, but only the cutting sense appeared in *Longman*), *integral* (a noun describing a mathematical function vs. the 'necessary part' sense in *Longman*), or *harmonic* (an adjective describing a type of function or series, but only the musical sense was given in the dictionary). For TIME the domain specific senses were cases like: *die* (a German article, but only defined in the noun or verb senses), *crimp* (as in 'a hindrance'; *Longman* only defines it as a verb), and *orient* (in the sense of finding a direction, but only defined in the Asian sense).

The experiment not only turned up new word meanings, it also identified several cases in which the dictionary was in error (the category *Longman error*). Many of these were differences between the machine-readable version of the dictionary and the printed version; these were cases that were caught by the proof reader when the printed dictionary was prepared (e.g., *majestic* defined as a noun, or *comfortable* as a verb). This illustrates that part-of-speech differences can not only be useful for identifying new word meanings, they can also be be an aid to proofreading during dictionary construction.

4.2 Morphology

Morphological gaps were determined by analyzing the 106 suffixes listed in *Longman*. The terms that ended with each suffix were extracted from each test collection, and the most frequent suffixes were identified. This data was used to build a morphological analyzer which would reduce a variant form to a word found in the dictionary. However, some root forms were not found in the dictionary, and there is a tradeoff between always finding the right root, and being flexible. For example, *capacitor* was not found in the dictionary, but we would like to recognize that it is related to *capacitance*. How do we know that *capacitor* is the correct root? If we are too flexible, we can end up reducing *digitize* to *digit*, and *factorial* to *factory* (in analogy to *matrimonial* being related to *matrimony*). Our analysis indicated that some endings were highly productive, and could be safely removed even though the root was not in the dictionary. These were: -ness, -ism, and -ly. The endings that were found to be very common combinations were also used to remove endings even if the root was not found. For example, -ization was always reduced to -ize.

Another way in which gaps were detected was to make use of subject codes. For example, in the NPL collection the word *ion* is always related to *ionic*, but *ionic* is defined in Longman as a type of Greek architecture. We would like to be able to recognize when the sense mentioned in the dictionary is not the same as the one in the text. This can be done by

using the morphological analyzer to recognize that *ionic* is a possible variant of *ion*, and then determining the dominant subject-code of the document (the dominant subject for a document is determined by looking up the subject codes for each word in the document; the subject code that occurs most often is the dominant subject code). If we have a possible variant, and the subject-code for the root form is the same as for the document in which it appears, that increases the likelihood that the possible variant is in fact correct. We tested this on the NPL collection, but found that it depends on what is considered the predominant subject. The dominant code for NPL is *science*, but more specifically it is *physics*. The *science* code occurs fairly often, and was found to cause too many false positives. That is, too many 'possible variants' were identified that were not actual variants. If a specialized code is used instead, most of the false positives do not occur. There were only 9 instances, however, in which the root was related to a variant whose meaning was not found in the dictionary. More work is needed with the other collections before this method can be considered reliable.

4.3 Subject Codes

The two previous sections were concerned with augmenting the dictionary using information from a corpus. In this section we will describe two experiments aimed at augmenting the dictionary by using the dictionary itself. This will be done by exploiting redundant information, and by recognizing links between senses and attempting to transfer information between them.

The machine-readable version of the *Longman Dictionary* includes subject codes associated with approximately 45% of its senses [Boguraev and Briscoe 87]. These codes are a two or four letter field, and indicate either a primary subject area, a primary and a secondary area, or a primary area and a specialization. For example, SI is the code for *science*, SIED is the code for *science* and a secondary code for *education*, and SIZP is the code for *science* and a specialized code for *physics*. The subject codes were not always assigned consistently, and in some cases the senses were assigned codes that are incorrect.

We tried two methods to detect senses that could have been assigned one of the codes:

1. Some definitions contain an indication of a domain within parentheses (e.g., *penalty* - '(in sports) a disadvantage given to a player or team for breaking a rule'). If the subject area indicated by the parenthetical (sports) did not match the subject code for that sense, it was identified as a candidate for assignment of that code. The word in parentheses will be referred to as a *domain label*.

2. Word overlap in the definitions of morphological variants. We will explain this in more detail below.

Comparison Result	Frequency
Code matched:	465 (75%)
Related code:	90 (15%)
Primary code missing:	13 (2%)
Specialized code missing:	13 (2%)
Secondary code missing:	2 (0%)
Codes were 'full':	4 (1%)
Errors:	11 (2%)
Compounds:	12 (2%)
Other:	10 (2%)

Table 4: Results of subject-code/domain-label comparison

To make use of the domain labels, all instances of '(in xyz)' were extracted from the text of the definitions. These were then sorted, and the list was examined to remove common instances that were not a reference to a subject area (e.g., '(in Britain)', '(in former times)', and '(in general)'). This resulted in a list of 757 items, which were processed semi-automatically to associate them with their corresponding subject code. Out of the 757 items, 620 were found to have a subject code that was an exact or close match. The 620 instances were then compared with the subject code associated with the sense for that instance. The results of this comparison are given in Table 4.

In 75% of the instances, the subject code was a match for the domain label. 'Related code' means that a closely related code was used instead of the one that matched the domain label. For example, *aeronautics* instead of *aerospace*, *science* instead of *engineering*, or *politics* instead of *military*. The next three lines refer to senses in which a primary, specialized, or secondary could have been assigned. 'Codes were full' means that a primary and secondary code had already been assigned, but that a third one (the domain label) was also applicable. 'Errors' means that the lexicographer used an incorrect code, such as PS (*psychology*) instead of SIZP (*physics*). 'Compounds' means that the subject code was a compound expression, such as *medicine and biology*, but the domain label was only one of them (this is an artifact of the matching routine, and they can also be grouped under 'Code matched').

A second method of finding subject-code gaps was also tried. In previous research we found that word-overlaps in the definitions of morphological variants are a good way of determining that the senses are related. If there is an overlap of two or more words, then the senses are strongly related more than 90% of the time[5] [Krovetz 93]. We identified the senses that were

Comparison Result	Frequency
transfer:	77 (37%)
connotation:	31 (15%)
mismatch:	26 (12%)
secondary:	22 (11%)
Longman error:	13 (6%)
Level mismatch:	5 (2%)
metaphor:	5 (2%)
unclear:	14 (7%)
other:	16 (8%)

Table 5: Subject code assignment via word overlap

strongly linked, and determined when they differed in their subject codes. These pairs were then examined manually to determine if the subject code could be assigned.

For the moment we have only examined the pairs for words beginning with the letters A, B, and C. A breakdown of the results is given in Table 5. There are 209 pairs, and 37% constitute clear cases for assigning the code. For some senses, there are differences in connotation. For example, *abstain* can refer to drinking or voting, and therefore has subject codes *beverages* and *politics*. The variant, *abstemious*, however, only has the connotation of abstaining from drinking or food. In contrast, the variant *abstention* only has the connotation of politics.

'Mismatch' refers to cases where the algorithm failed to identify a related sense. 'Secondary' means that a secondary code can transfer over, but not a primary one. 'Longman error' means that the code assigned by the lexicographer is incorrect. 'Level mismatch' refers to cases resulting from the way the subject codes are structured. For example, *sports* and *net games* are both primary codes. There are many cases in which a code would probably be better as a specialization.

Finally, we note that there is a potential for extending the *Longman* subject-codes with information acquired from a corpus. In our initial examination of the lexicons (see Table 2), we found that hyphenated words can provide a very good characterization of the subject matter of a corpus. For example, the most frequent hyphenated forms for the different test collections are: *time-sharing*, *context-free*, *on-line*, and *real-time* for CACM, *sino-soviet*, *anti-communist*, *left-wing*, and *cease-fire* for TIME, and *third-party*, *three-judge*, and *cross-examination* for WEST (unfortunately, almost all punctuation in the NPL collection was

omitted when the collection was created). In conjunction with the existing *Longman* codes, these hyphenated words can help to confirm the characterization, and refine it even further.

5 Conclusion

While it is recognized that dictionaries must be supplemented with information from corpora, relatively little quantitative data is available about the extent of the gap. We conducted experiments to determine the coverage provided by a learner's dictionary (*Longman*), and how many additional words would be found by using a larger dictionary (*Collins*). We then explored various methods for identifying missing information in lexical entries.

The experiments show that the coverage of the Longman dictionary is very good; only a small number of the words not found in it are found in the larger Collins dictionary. The words that *are* found are typically technical words, compounds, prefixed forms, and abbreviations.

We explored several methods to find gaps in the dictionary, i.e., places where information associated with the senses was incomplete. These included using a stochastic tagger to identify part-of-speech, a morphological analyzer to determine variants not specified, and exploiting information within the dictionary to identify missing subject codes. We were able to successfully identify gaps for each type of missing information, but it was not possible to prevent a significant number of false positives. Problems were caused by differences in sense connotation, reliability of subject code assignment, and reliability of word overlap for identifying related senses. Surprisingly, even though stochastic taggers are reported to have a high accuracy rate, tagging error was a significant problem; many of the false positives for part-of-speech were a result of tagging error.

While the error rates encountered are too high to allow for full-automatic augmentation of the lexicon, these methods can be used to help the lexicographer identify new words and word-senses. There are also questions about how much impact these gaps have on particular applications. We are currently conducting experiments on word sense disambiguation and information retrieval, and the impact of these gaps will be reported in a future paper.

Notes

1. *Longman* also includes about 7,000 phrasal headwords, such as *hot line*, and *line printer*. We wanted to avoid the issue of phrases for the moment, so this part of the analysis has only been done with individual words.

2. These were compiled from lists of first and last names, and lists of locations; it is intended as a means of capturing common proper nouns. Other proper nouns will be captured by the 'capitalized words' category.

3. A stochastic tagger uses statistical information to assign a part-of-speech tag to a word in context. These taggers typically combine lexical probabilities with statistics about the likelihood of various tag sequences.

4. These pairs did not include differences involving words tagged as proper nouns. Although they sometimes reflected meaning differences (e.g., the names of programming languages: BASIC, BLISS, COMPASS, GASP, JOVIAL, LISP), we found that too many false positives were generated due to capitalized words occurring in the titles of documents.

5. The overlap does not include closed class words, and reduces all inflected forms in the definitions to their root forms.

Acknowledgments

This work was supported by the Center for Intelligent Information Retrieval at the University of Massachusetts. I am grateful to Scott Anderson and John Broglio for their comments on a draft version of this paper.

References

[Belmore 88] Belmore N, "The Use of Tagged Corpora in Defining Informationally Relevant Word Classes", in *Corpus Linguistics: Hard and Soft*, J Aarts and W Meijs (eds), Rodopi Press, 1988

[Boguraev and Briscoe 87] Boguraev B and Briscoe T, *Computational Lexicography for Natural Language Processing*, Longman, 1987

[Church 88] Church K, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," *Proceedings of the 2nd Conference on Applied Natural Language Processing,* pp. 136–143, 1988.

[Krovetz 93] Krovetz R, "Viewing Morphology as an Inference Process", *Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 191-202, 1993

[Proctor 78] Proctor P, *Longman Dictionary of Contemporary English*, Longman, 1978.