Oxford Dictionary of English: Current Developments

James McCracken
Oxford University Press
mccrackj@oup.co.uk

Abstract

This research note describes the early stages of a project to enhance a monolingual English dictionary database as a resource for computational applications. It considers some of the issues involved in deriving formal lexical data from a natural-language dictionary.

1 Introduction

The goal of the project is to enhance the database of the *Oxford Dictionary of English* (a forthcoming new edition of the 1998 *New Oxford Dictionary of English*) so that it contains not only the original dictionary content but also additional sets of data formalizing, codifying, and supplementing this content. This will allow the dictionary to be exploited effectively as a resource for computational applications.

The *Oxford Dictionary of English* (ODE) is a high-level dictionary intended for fluent English speakers (especially native speakers) rather than for learners. Hence its coverage is very extensive, and definitional detail is very rich. By the same token, however, a certain level of knowledge is assumed on the part of the reader, so not everything is spelled out explicitly. For example, ODE frequently omits morphology and variation which is either regular or inferable from related words. Entry structure and defining style, while mostly conforming broadly to a small set of basic patterns and formulae, may often be more concerned with detail and accuracy than with simplicity of explanation. Such features make the ODE content relatively difficult to convert into comprehensive and formalized data. Nevertheless, the richness of the ODE text, particularly in the frequent use of example sentences, provides a wealth of cues and clues which can help to control the generation of more formal lexical data.

A basic principle of this work is that the enhanced data should always be predicated on the original dictionary content, and not the other way round. There has been no attempt to alter the original content in order to facilitate the generation of formal data. The enhanced data is intended primarily to constitute a formalism which closely reflects, summarizes, or extrapolates from the existing dictionary content.

The following sections list some of the data types that are currently in progress:

2 Morphology

A fundamental building block for formal lexical data is the creation of a complete morphological formalism (verb inflections, noun plurals, etc.) covering all lemmas (headwords, derivatives, and compounds) and their variant forms, and encoding relationships between them. This is being done largely automatically, assuming regular patterns as a default but collecting and acting on anything in the entry which may indicate exceptions (explicit grammatical information, example sentences, pointers to other entries, etc.).

The original intention was to generate a morphological formalism which reflected whatever was stated or implied by the original dictionary content. Hence pre-existing morphological lexicons were not used except when an ambiguous case needed to be resolved. As far as possible, issues relating to the morphology of a word were to be handled by collecting evidence internal to its dictionary entry.

However, it became apparent that there were some key areas where this approach would fall short. For example, there are often no conclusive indicators as to whether or not a noun may be plu-

ralized, or whether or not an adjective may take a comparative or superlative. In such cases, any available clues are collected from the entry but are then weighted by testing possible forms against a corpus.

3 Idioms and other phrases

Phrases and phrasal verbs are generally lemmatized in an 'idealized' form which may not represent actual occurrences. Variation and alternative wording is embedded parenthetically in the lemma:

(as) nice (or *sweet) as pie*

Objects, pronouns, etc., which may form part of the phrase are indicated in the lemma by words such as 'someone', 'something', 'one':

*twist (*or *wind* or *wrap) someone around one's little finger*

In order to be able to match such phrases to real-world occurrences, each dictionary lemma was extended as a series of strings which enumerate each possible variant and codify how pronouns, noun phrases, etc., may be interpolated. Each occurrence of a verb in these strings is linked to the morphological data in the verb's own entry, to ensure that inflected forms of a phrase (e.g. 'she had him wrapped around her little finger') can be identified.

4 Semantic classification

We are seeking to classify all noun senses in the dictionary according to a semantic taxonomy, loosely inspired by the Princeton WordNet project. Initially, a relatively small number of senses were classified manually. Statistical data was then generated by examining the definitions of these senses. This established a definitional 'profile' for each classification, which was then used to automatically classify further senses. Applied iteratively, this process succeeded in classifying all noun senses in a relatively coarse-grained way, and is now being used to further refine the granularity of the taxonomy and to resolve anomalies.

Definitional profiling here involves two elements:

The first element is the identification of the 'key term' in the definition. This is the most significant noun in the definition – not a rigorously defined concept, but one which has proved pragmatically effective. It is not always coterminous with the genus term; for example, in a definition beginning 'a morsel of food which…', the 'key term' is taken to be *food* rather than *morsel*.

The second element is a scoring of all the other meaningful vocabulary in the definition (i.e. ignoring articles, conjunctions, etc.). A simple weighting scheme is used to give slightly more importance to words at the beginning of a definition (e.g. a modifier of the key term) than to words at the end.

These two elements are then assigned mutual information scores in relation to each possible classification, and the two MI scores are combined in order to give an overall score. This overall score is taken to be a measure of how 'typical' a given definition would be for each possible classification. This enables one very readily to rank and group all the senses for a given classification, thus exposing misclassifications or points where a classification needs to be broken down into subcategories.

The semantic taxonomy currently has about 1250 'nodes' (each representing a classification category) on up to 10 levels. The dictionary contains 95,000 defined noun senses in total, so there are on average 76 senses per node. However, this average disguises the fact that there are a small number of nodes which classify significantly larger sets of senses. Further subcategorization of large sets is desirable in principle, but is not considered a priority in all cases. For example, there are several hundred senses classified simply as *tree*; the effort involved in subcategorizing these into various tree species is unlikely to pay dividends in terms of value for normal NLP applications. A pragmatic approach is therefore to deprioritize work on homogeneous sets (sets where the range of 'typicality' scores for each sense is relatively small), more or less irrespective of set size.

Hence the goal is not achieve granularity on the order of WordNet's 'synset' (a set in which all terms are synonymous, and hence are rarely more than four or five in number) but rather a somewhat more coarse-grained 'similarset' in which every sense is similar enough to support general-purpose word-sense disambiguation, document retrieval, and other standard NLP tasks. At this level, auto-

matic analysis and grading of definitions is proving highly productive in establishing classification schemes and in monitoring consistency, although extensive supervision and manual correction is still required.

It should be noted that a significant number of nouns and noun senses in ODE do not have definitions and are therefore opaque to such processes. Firstly, some senses cross-refer to other definitions; secondly, derivatives are treated in ODE as undefined subentries. Classification of these will be deferred until classification of all defined senses is complete. It should then be possible to classify most of the remainder semi-automatically, by combining an analysis of word formation with an analysis of target or 'parent' senses.

5 Domain indicators

Using a set of about 200 subject areas (*biochemistry*, *soccer*, *architecture*, *astronomy*, etc.), all relevant senses and lemmas in ODE are being populated with markers indicating the subject domain o which they relate. It is anticipated that this will support the extraction of specialist lexicons, and will allow the ODE database to function as a resource for document classification and similar applications.

As with semantic classification above, a number of domain indicators were assigned manually, and these were then used iteratively to seed assignment of further indicators to statistically similar definitions. Automatic assignment is a little more straightforward and robust here, since most of the time the occurrence of strongly-typed vocabulary will be a sufficient cue, and there is little reason to identify a key term or otherwise parse the definition.

Similarly, assignment to undefined items (e.g. derivatives) is simpler, since for most two- or three-sense entries a derivative can simply inherit any domain indicators of the senses of its 'parent' entry. For longer entries this process has to be checked manually, since the derivative may not relate to all the senses of the parent.

Currently, about 72,000 of a total 206,000 senses and lemmas have been assigned domain indicators. There is no clearly-defined cut-off point for iterations of the automatic assignment process; each iteration will continue to capture senses which are less and less strongly related to the domain. Beyond a certain point, the relationship will become too tenuous to be of much use in most contexts; but that point will differ for each subject field (and for each context). Hence a further objective is to implement a 'points' system which not only classifies a sense by domain but also scores its relevance to that domain.

6 Collocates for senses

We are currently exploring methods to automatically determine key collocates for each sense of multi-sense entries, to assist in applications involving word-sense disambiguation. Since collocates were not given explicitly in the original dictionary content of ODE, the task involves examining all available elements of a sense for clues which may point to collocational patterns.

The most fruitful areas in this respect are firstly definition patterns, and secondly example sentences.

Definition patterns are best illustrated by verbs, where likely subjects and or objects are often indicated in parentheses:

fly: (of a bird, bat, or insect) move through the air…
impound: (of a dam) hold back (water)…

The terms in parentheses can be collected as possible collocates, and in some cases can be used as seeds for the generation of longer lists (by exploiting the semantic classifications described in section 3 above). Similar constructions are often found in adjective definitions. For other parts of speech (e.g. nouns), and for definitions which happen not to use the parenthetic style, inference of likely collocates from definition text is a less straightforward process; however, by identifying a set of characteristic constructions it is possible to define search patterns that will locate collocate-like elements in a large number of definitions. The defining style in ODE is regular enough to support this approach with some success.

Some notable 'blind spots' have emerged, often reflecting ODE's original editorial agenda; for example, the defining style used for verbs often makes it hard to determine automatically whether a sense is transitive or intransitive.

Example sentences can be useful sources since they were chosen principally for their typicality,

and are therefore very likely to contain one or more high-scoring collocates for a given sense. The key problem is to identify automatically which words in the sentence represent collocates, as opposed to those words which are merely incidental. Syntactic patterns can help here; if looking for collocates for a noun, for example, it makes sense to collect any modifiers of the word in question, and any words participating in prepositional constructions. Thus if a sense of the entry for *breach* has the example sentence

She was guilty of a breach of trust.

then some simple parsing and pattern-matching can collect *guilty* and *trust* as possible collocates.

However, it will be apparent from this that examination of the content of a sense can do no more than build up lists of *candidate* collocates – a number of which will be genuinely high-scoring collocates, but others of which may be more or less arbitrary consequences of an editorial decision. The second step will therefore be to build into the process a means of testing each candidate against a corpus-based list of collocates, in order to eliminate the arbitrary items and to extend the list that remains.

7 Conclusion

In order for a non-formalized, natural-language dictionary like ODE to become properly accessible to computational processing, the dictionary content must be positioned within a formalism which explicitly enumerates and classifies all the information that the dictionary content itself merely assumes, implies, or refers to. Such a system can then serve as a means of entry to the original dictionary content, enabling a software application to quickly and reliably locate relevant material, and guiding interpretation.

The process of automatically generating such a formalism by examining the original dictionary content requires a great deal of manual supervision and ad hoc correction at all stages. Nevertheless, the process demonstrates the richness of a large natural-language dictionary in providing cues and flagging exceptions. The stylistic regularity of a dictionary like ODE supports the enumeration of a finite (albeit large) list of structures and patterns which can be matched against a given entry or entry element in order to classify it, mine it for pertinent information, and note instances which may be anomalous.

The formal lexical data is being built up alongside the original dictionary content in a single integrated database. This arrangement supports a broad range of possible uses. Elements of the formal data can be used on their own, ignoring the original dictionary content. More interestingly, the formal data can be used in conjunction with the original dictionary content, enabling an application to exploit the rich detail of natural-language lexicography while using the formalism to orient itself reliably. The formal data can then be regarded not so much as a stripped-down counterpart to the main dictionary content, but more as a bridge across which applications can productively access that content.

Acknowledgements

I would like to thank Adam Kilgarriff of ITRI, Brighton, and Ken Litkowski of CL Research, who have been instrumental in both devising and implementing significant parts of the work outlined above.

References

Christiane Fellbaum and George Miller. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, Mass.

Judy Pearsall. 1998. *The New Oxford Dictionary of English*. Oxford University Press, Oxford, UK.