Rebuilding the *Oxford Dictionary of English* as a semantic network

James McCRACKEN
Oxford University Press
Great Clarendon Street
Oxford OX2 6DP, UK
james.mccracken@oup.com

Abstract

This paper describes a project to develop a lexicon for use both as an electronic dictionary and as a database for a range of NLP tasks. It proposes that a lexicon for such open-ended application may be derived from a human-user dictionary, retaining and enhancing the richness of its editorial content but abandoning its entry-list structure in favour of networks of relationships between discrete lexical objects, where each object represents a discrete lexeme-meaning unit.

1 Introduction

Dictionaries intended for human users typically sacrifice strict formalism, structure, and consistency in the interests of space and readability. Furthermore, a dictionary may assume a certain level of knowledge on the part of the reader, and so need not always be exhaustively explicit about all aspects of a given lexeme. On the other hand, traditional high-level lexicography is valuable for the rich variety of detailed, complex information integrated in each entry, usually with very low error rates. Clearly, then, there are both benefits and problems in attempting to mine or query such a dictionary computationally.[1] The project described here attempts to retain the rich editorial content of a dictionary but to reorganize it in structured networks that are more readily processed and traversed - in other words, to preserve the benefits while smoothing out some of the problems.

2 Method

The major steps in this process were as follows:

1. Decomposition of the dictionary as a flat list of entries, to be replaced by a set of lexical objects (each corresponding roughly to a single lexeme-meaning pair in the original dictionary, and minimally retaining all the text of that meaning) (section 4 below).
2. Population of each lexical object with a complete set of morphological and syntactic data (section 5 below).
3. Classification connecting lexical objects to each other in semantic hierarchies and networks of domain relationships (section 6 below).

3 Choice of dictionary

The dictionary selected was the *Oxford Dictionary of English* (2003) (ODE). This is a relatively high-level dictionary, intended for fluent English speakers rather than learners. Accordingly, it assumes a body of general knowledge and a degree of semantic and grammatical competence on the part of the user: not everything is made explicit in the text.[2] Computational analysis of the text must be designed to identify and aggregate cues, and to employ rules for inheriting information from entry or branch level down to sense level, in order to generate the right data to construct each lexical object (see section 4 below).

On the other hand, ODE has some key features that make it particularly appropriate for enhancement a as comprehensive electronic lexical database, usable both as a dictionary and as a resource for exploitation in NLP applications:

- even-handed treatment of both British and American spellings, plus extensive coverage of other World English;
- coverage of proper-name items (places, people, events, etc.);
- relatively codified editorial style, facilitating computational evaluation of definitions;
- detailed and codified indication of syntactic characteristics, sense by sense;
- example sentences linked to individual senses, providing cues about grammatical behaviour and collocation.

Furthermore, ODE is very much a corpus-based dictionary: all key editorial decisions - including

[1] See Ide and Veronis 1993 for a good account of this tension.

[2] An example of this in ODE is the high number of run-on derivatives, which are undefined on the assumption that the user will be able to infer the sense morphologically.

orthography, sense distinctions, choice of example sentences, and syntactic information - are motivated primarily by corpus evidence. This means that a formalized encoding of the information presented in the dictionary is likely to generate robust and comprehensive resources for NLP tasks dealing with arbitrary real-world text.

4 Data structure and lexical objects

The ODE data structure dispenses with the idea of the *entry* as the dictionary's basic unit. This move reflects the proposition that the traditional dictionary entry is an artefact of human-readable dictionaries, and expresses a predominantly etymological view of the language. The dictionary entry brings together lexical elements (definitions, parts of speech, phrases, etc.) which may have little relationship to each other from the point of view of language use (meaning, domain, frequency, collocation, etc.). The entry is therefore an inappropriate unit for computer processing.

Instead, the ODE data structure is built as a series of discrete lexical objects which correspond roughly to individual dictionary senses or to senses of embedded lemmas (phrases, compounds, derivatives, etc.). These lexical objects function as packets of data which exist independently of each other, each containing all relevant information about meaning, morphology, lexical function, semantic category, etc. Hence every sense may be queried, extracted, or manipulated as an independent object without any reference to the entry in which it was originally situated.

Although search results, etc., may still (typically) be represented to the user in the form of traditional dictionary entries, from a functional point of view the entry is rendered redundant as a data object, and for almost all purposes is ignored when handling search queries and other processing.

The choice of lexeme-meaning pair rather than entry as the basic lexical object allows a much more detailed and exhaustive specification of the way the language functions on a sense-by-sense basis. Each object typically includes (minimally) a definition, one or more example sentences, a full specification of morphology and syntactic behaviour (see section 5 below), and a range of classifiers (see section 6 below). This creates some redundancy: for example, an entry with numerous senses may generate several lexical objects which repeat the same morphology and syntactic behaviour. However, this is more than offset by the advantages: by integrating data into self-contained lexical objects, the structure (a) supports robust and positive Boolean operations and (b) opens up the lexicon for reconfiguration according to any new type of lexical or semantic relation.

5 Morphological and phrasal specification

Every lexical object in ODE provides a complete specification of all the lexical forms relevant to its sense. This includes not only the morphology of the lexemes themselves, but also structured data about their syntactic roles, variant spellings, British and US spellings, and alternative forms. This is true not only for single-word lexemes but also for abbreviations, initialisms, affixes, contractions, compounds, phrases, and proper names. ODE thus provides facility for look-up of real-world lexical forms in context, including:

As mentioned above, such data was generated for every lexical object (sense). For some polysemous headwords this may mean multiple repetitions of morphological listings. More typically, however, listings vary from sense to sense. One common variable, for example, is that for a given noun headword, some lexical objects may include a plural while others do not. Similarly, variant spellings may often be associated with some but not all lexical objects. More pertinently perhaps, phrasal patterns associated with the lexeme will vary significantly from sense to sense, particularly for verbs.

It should be noted that this formal variability can play an important role in the early stages of disambiguation tasks; the particular form of a lexeme or extended phrase may often allow groups of senses to be reliably discounted or at least downgraded as possible meanings.[3] This simplifies the set of candidate meanings subsequently presented to semantic analysis algorithms.

Hence it was judged necessary to develop distinctive morphological and phrasal specifications for each lexical object. Principally this was done algorithmically: the dictionary text associated with each lexical object was examined for clues not only in explicit syntax information but also in definition patterns and especially in example sentences. In default of any strong pointers emerging from this process, forms were produced by a simple morphological generator.

However, it became apparent that many formal variations were not cued in any way by the dictionary text. As a simple example, if an adjective is gradable then this is always indicated

[3] John Sinclair has proposed that 'Every distinction in meaning is associated with a distinction in form' (quoted in Hanks 2003, p. 61). Although doubtful about this as a universal rule, I agree that for a given lexeme one may often link major categories of meaning to typical formal characteristics.

at headword level in the dictionary, but there is nothing to indicate which *senses* are gradable and which are not. Because of limitations like this, a significant stage of post-processing manual correction was necessary.

6 Classification

Each lexical object includes a set of classifiers which are used to position the object's definition in semantic taxonomies (structures of superordination/subordination) and in subject areas (domains).

This may be used in two principal ways. Firstly, it allows the dictionary to be navigated along associative or thesaurus-like routes: it structures the lexicon's set of objects into a network of connections between meanings. Secondly, it allows calculations of semantic distance between lexical objects, supporting NLP tasks relating to sense disambiguation, context-sensitive look-up, and document classification.[4]

6.1 Semantic taxonomy

All lexical objects representing noun meanings (about 100,000) have been structured in a semantic taxonomy. This represents a substantial majority of all objects; the smaller sets of verbs and adjectives will be similarly structured as the next stage of the project.

The taxonomy's high-level nodes (major classifications) were originally constructed to follow the file structure of WordNet nouns. Although the ODE taxonomy has subsequently diverged in detail, mappings to WordNet have been retained. At the early stages of the project, a relatively small number of noun objects were classified manually. The definitions contained in these objects were then used as training data to establish a definitional 'profile' for each high-level node. Definitional profiles were used to automatically classify further objects. Applied iteratively (with manual correction at each stage), this process succeeded in classifying all noun objects in a relatively coarse-grained way.

Beyond this stage, the statistical data which had previously been integrated to build definition profiles was instead decomposed to help identify ways in which high-level nodes could be subdivided to develop greater granularity.

Definitional profiling here involves two elements:

1. Identification of the 'key term' in the definition. This is the most significant noun in the definition. It is not always coterminous with the genus term; for example, in a definition beginning 'a morsel of food which…', the 'key term' is calculated to be *food* rather than *morsel*.
2. Stemming and scoring of all other meaningful lexemes in the definition (ignoring articles, conjunctions, etc.). A simple weighting scheme is used to give more importance to lexemes at the beginning of a definition (e.g. a modifier of the key term) than to lexemes 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 a given classification. This enables one very readily to rank all the lexical objects attached to a given node, and to identify other objects which are candidates for that node.

The semantic taxonomy currently has about 2200 nodes on up to 12 levels - on average, 46 objects per node. However, this average disguises the fact that there are a small number of nodes which classify significantly larger sets of objects. Further subcategorization of large sets is desirable in principle, but is not considered a priority in all cases: subcategorization of a given set is deprioritized if the set if relatively homogeneous, i.e. if the distribution of 'typicality' scores for each object is relatively small.[5]

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). Instead, granularity is based on a more thesaurus-like measure of parity between objects in a set.

6.2 Domains

As with semantic classification, 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 necessity

[4] Metrics for using taxonomies to calculate semantic similarity are discussed in Leacock and Chodorow 1998. An implementation of ODE as an electronic dictionary uses such metrics in the automatic glossing of its own text.

[5] For example, the *tree* set is several times larger than average, but since tree definitions have similar profiles, the set produces a high homogeneity score. This accords loosely with the intuition that for most NLP purposes, there is little value in making semantic distinctions between different species of tree.

to identify a key term or otherwise to parse the definition.

Not every lexical object is domain-specific: currently 49% of all objects have been assigned at least one domain marker. Each iteration of the assignment process will continue to capture objects 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 grading scheme which not only classifies an object by domain but also scores its relevance to that domain.

Currently, a set of about 220 domains are used, both technical (*Pharmaceutics, Oceanography* ,) and thematic (*Crime, Sex*). These were originally planned to be organized in a Dewey-like taxonomy. However, this was found to be unworkable, since within the lexicon domains emerged as a series of overlapping fields rather than as a hierarchy. Hence the domains have now been reorganized not taxonomically but rather as a network of multiple connections. Within the network, each edge connecting two domains represents the strength of association between those domains.[6]

7 Work in progress

Ongoing work focuses mainly on the cumulative population of lexical objects with additional data fields, in particular collocations and frequency measures.

Additionally, further classifier types are being trialled in order to define further relations between lexical objects. These include linguistic relations such as antonymy and root derivation, and real-world relations such as geographical region.

8 Conclusion

The strategy used in constructing the ODE electronic database was to preserve the full editorial content of a human-user dictionary but to rebuild its structure, replacing the entry-list paradigm with a manifold network of relations between meanings.

The key benefit of this approach is in the versatility of the database. Lexical objects may be reassembled into entries for display, so ODE can still function as an electronic human-user dictionary, albeit one that takes advantage of novel search and navigation features. Additionally, ODE is directly usable in a number of non-dictionary applications. These include context-sensitive spellchecking, tagging and parsing, document categorization, and context-sensitive document glossing.

Feedback from such applications is being monitored not only to critically examine and correct the source data, but also to examine the source dictionary itself: because much of the formal and classificatory data is generated algorithmically from analysis of the source editorial content of each lexical object (definition, etc.), anomalies emerging in that data can often be traced back to anomalies in the editorial content (e.g. inconsistencies in defining style).

The ODE project is therefore in part an attempt to bridge the distinction between human-user dictionaries and WordNet-like associative electronic lexicons. By deriving the one from the other, it invites applications to navigate and mine the rich lexicographic content of the original dictionary by means of a new set of structured relations and frameworks.

Acknowledgements

I would like to thank Adam Kilgarriff of ITRI, Brighton, who has played a key role in advising on many aspects of the work described here.

References

Hanks, P. 2003. 'Lexicography'. In R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics* . Oxford: Oxford University Press.

Ide, N. and J. Veronis, 1993. 'Extracting knowledge bases from machine-readable dictionaries: have we wasted our time?'. in *KB&KS Workshop* . Tokyo.

Leacock, C. and M. Chodorow. 1998. 'Combining local context and WordNet similarity for word sense identification'. In C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database* . Cambridge, Mass.: MIT Press.

Rigau, G., 1994. 'An experiment on automatic semantic tagging of dictionary senses'. In *Proceedings of the workshop 'The Future of Dictionary'* . Aix-les-Bains, France.

Soanes, C. and A. Stevenson. 2003. *Oxford Dictionary of Eng lish*. Oxford: Oxford University Press.

[6] Strength of association from Domain_A to Domain_B is determined internally to ODE, by calculating the proportion of (a) lexemes and (b) semantic sets in which both domains appear, as opposed to those in which only Domain_A appears. Strength of association is not mutual.