Evaluating Lexical Resources for A Semantic Tagger

Scott S. L. Piao[1], Paul Rayson[2], Dawn Archer[1], Tony McEnery[1]
[1]Department of Linguistics and MEL
[2]Computing Department
Lancaster University
Lancaster LA1 4YT
United Kingdom
{s.piao@lancaster.ac.uk;paul@comp.lancs.ac.uk;d.archer@lancaster.ac.uk; amcenery@lancaster.ac.uk }

Abstract

Semantic lexical resources play an important part in both linguistic study and natural language engineering. In Lancaster, a large semantic lexical resource has been built over the past 14 years, which provides a knowledge base for the USAS semantic tagger. Capturing semantic lexicological theory and empirical lexical usage information extracted from corpora, the Lancaster semantic lexicon provides a valuable resource for the corpus research and NLP community. In this paper, we evaluate the lexical coverage of the semantic lexicon both in terms of genres and time periods. We conducted the evaluation on test corpora including the BNC sampler, the METER Corpus of law/court journalism reports and some corpora of Newsbooks, prose and fictional works published between 17th and 19th centuries. In the evaluation, the semantic lexicon achieved a lexical coverage of 98.49% on the BNC sampler, 95.38% on the METER Corpus and 92.76% -- 97.29% on the historical data. Our evaluation reveals that the Lancaster semantic lexicon has a remarkably high lexical coverage on modern English lexicon, but needs expansion with domain-specific terms and historical words. Our evaluation also shows that, in order to make claims about the lexical coverage of annotation systems as well as to render them 'future proof', we need to evaluate their potential both synchronically and diachronically across genres.

1. Introduction

Lexical resources play an important part in both linguistic study and natural language engineering. Over the past decade, in particular, large semantic lexicons, such as WordNet (Fellbaum, 1998), EuroWordNet (Vossen, 1998), HowNet (http://www.keenage.com), etc. have been built and applied to various tasks.

During the same period of time, another large semantic lexical resource has been built in Lancaster University, as a knowledge base for an English semantic tagger named USAS (Rayson and Wilson 1996; Piao *et al*. 2003). Employing a semantic annotation scheme, this lexicon links English lexemes and multiword expressions to their potential semantic categories, which are disambiguated according to their context in actual discourse.

In this paper, we present our evaluation work on the lexical coverage of the semantic lexicon of the Lancaster semantic tagger. During the evaluation, we examined the system's lexical coverage in both modern general English and a narrow-domain English corpus. We also investigated how the time periods affect the lexical coverage of our semantic lexicon. As this paper will show, our evaluation suggests that the optimal way of evaluating lexical resources is to conduct it over multiple genres and various time periods, using a large representative corpus or several domain-specific corpora.

2. Lancaster Semantic Lexicon

As mentioned earlier, the Lancaster semantic lexicon has been developed as a semantic lexical knowledge database for a semantic tagger. It consists of two main parts: a single word sub-lexicon and a multi-word expression (MWE) sub-lexicon. Currently it contains over 42,300 single word entries and over 18,400 multi-word expression entries.

In the single word sub-lexicon, each entry maps a word, together with its POS category[1], to its potential semantic categories. For example, the word "iron" is mapped to the category of {object/substance and material} when it is used as a noun, and to the category of {cleaning and personal care} when it is used as a verb. When provided with context, these candidate categories can be disambiguated.

The entries in the MWE lexicon have similar structures as the single word counterpart but the key words are replaced by MWEs. Here, the constituent words of each MWE are considered as a single semantic entity, and thus mapped to semantic category/ies together. For example, the MWE "life expectancy" is mapped to the categories of {time/age} and {expect}.

In addition, to account for MWEs of similar structures with the same entry, many MWEs are transcribed as templates using a simplified form of regular expression. For example, the template {*ing_NN1 machine*_NN*} represents a set of MWEs including "washing machine/s", "vending machine/s", etc. As the result, the MWE lexicon covers many more MWEs than the number of individual entries. Furthermore, the templates also capture discontinuous MWEs.

The Lancaster semantic taxonomy was initially based on Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981), but has undergone a series of expansion and improvements. Currently it contains 21 major discourse fields that expand, in turn, into 232 categories (for further details

1 In the Lancaster semantic lexicon, the C7 POS tagset is used to encode POS information.

of the semantic taxonomy, see website: http://www.comp.lancs.ac.uk/ucrel/usas/).

The Lancaster semantic lexicon is presently being expanded and improved as part of the Benedict Project (EU project IST-2001-34237). In the following sections, we describe our evaluation of the lexical coverage of the current semantic lexicon.

3. Test Corpora

Our aim, in this evaluation, was to evaluate the general lexical coverage potential of the USAS semantic tagger as well as to investigate how factors like genre, domain and historical period affect the lexical coverage. Accordingly, we selected test corpora that reflect a variety of genres/domains and time periods.

First, we selected the BNC sampler corpus to represent general modern English in this evaluation. Containing about two million words, this corpus consists of similarly sized texts from various genres. In addition, it contains two equally sized written and spoken sections. With such diversity of its contents, it has been considered to be highly representative of modern English[2]. We also used the written and spoken parts of the BNC sampler separately as a means of assessing the lexical coverage in written and spoken genres.

Next, we chose the Meter Corpus (Gaizasukas *et al*. 2001) as our narrow domain test corpus for estimating the lexical coverage in specific domains. The corpus contains journalistic reports from the UK Press Association (PA) newswire service and similar reports from nine UK mainstream newspapers. However, [3] we only used the newspaper reports on law/court stories from the corpus for our evaluation, as we assumed that, with the size of 241,311 words and with the content constrained to law and court events, this data provides an ideal means of testing the lexical coverage on a narrow domain.

With regard to the diachronic factor concerning the lexical coverage, we drew test corpora from two sources. The first test corpus (of 61, 065 words) was taken from the Lancaster Newsbooks Corpus (LNC) [4], a collection of English newsbooks published in the 17th century. In our evaluation, a section of it containing 61,065 words was selected. The second test corpus containing 6,544,342 words was taken from a collection of prose/fictional works from the 18th and 19th centuries.

As explained above, we selected the test corpora from a variety of sources to ensure that the result of our evaluation truly reflects the lexical coverage of our system in practical annotation tasks. Although the sources of our test corpora are not sufficiently broad to claim a complete representation of the English language, we believe that they are sufficiently diverse to gain an insight into the general lexical coverage of our tagging system.

4. Evaluation

Generally, there are two ways of evaluating lexical coverage, as Demetriou and Atwell (2001) put it:

- one that uses the number of distinct word forms in text ("vocabulary type" coverage). This answers the question "*how many of the different word types in language are covered by the system?*"; …
- one that uses the total number of words in text (the probability of finding a root for a word token – "real text token" coverage); this answers question "*how many of the word tokens in a text are expected to be covered by the system?*" …

We took the second approach in our evaluation, i.e. we used the number of tokens rather than the number of word types as the base number for our statistics. This is because our evaluation was conducted as a part of the test of performance of the USAS semantic tagger, and hence we focused on investigating the impact of the missing words in our lexicon on the processing of running texts in practical tagging tasks. For this particular evaluation, we assumed that the percentage of identified words in terms of tokens is more significant than that of word types.

We conducted the evaluation as follows. We first semantically tagged the test corpora, marking the words not found in our lexicon, then collected and counted these words to calculate the lexical coverage. The figures of mismatches that we report below include typos and other non-words. For reasons of space, we do not go into the details in this paper.

The first step of the evaluation involved an examination of the lexical coverage on modern English corpora, i.e. the BNC Sampler Corpus and METER Corpus were the test data. Table 1 shows the lexical coverage on each of these test corpora.

Test Corpus	Total Tokens	Unmatched Tokens	Lexical Coverage
BNC Sampler	1,956,171	29,517	98.49%
BNC Samp. Written sect.	970,532	23,407	97.59%
BNC Samp. Spoken sect.	985,639	6,110	99.39%
METER Corpus	241,311	11,143	95.38%

Table1: Lexical coverage of Lancaster semantic lexicon on modern English test corpora[5]

[2] For further details of BNC sampler, see website: http://www.natcorp.ox.ac.uk/getting/sampler.html.

[3] For further details of the METER Corpus, see website: http://www.dcs.shef.ac.uk/nlp/funded/meter.html.

[4] For further details of the LNC, see website http://www.ling.lancs.ac.uk/newsbooks.

[5] In this table, the BNC Sampler written and spoken sub-corpora are two sections of the BNC Sampler Corpus.

As shown in Table 1, our semantic lexicon obtained lexical coverage ranging between 99.39% and 95.38% on the test corpora. By and large, it obtained slightly better coverage on modern general language than on the domain specific corpora.

Our lexicon achieved an encouraging lexical coverage when applied to the BNC Sampler corpus (which, as we highlight above, is assumed to represent modern general English). Out of the total 1,956,171 tokens in the corpus, it failed to identify 29,517 tokens, resulting in a lexical coverage of 98.49%. Such a high lexical coverage is the result of continuous improvement of the lexicon over the past decade. It shows that the Lancaster semantic lexicon is capable of dealing with most general domains.

Next, in order to investigate the influence of written and spoken genres of English on the lexical coverage, we examined the lexical coverage on the written and spoken sections of the BNC Sampler corpus separately. As shown in Table 1, we found that the majority of the unmatched words were in the written section of the corpus. To elaborate, 23,407 out of the total 970,532 tokens were not matched in the written section, producing a coverage rate of 97.59%. In contrast, only 6,110 out of the total 985,639 tokens in the spoken section were unmatched, producing a coverage rate of 99.39%. As the vocabulary of spoken language tends to be smaller than that of written language, such a result is hardly surprising. In addition, more common words tend to be used in spoken language than in the written texts.

The Lancaster semantic lexicon has been built to mainly deal with general English, collected from sources like the balanced BNC corpus. We therefore anticipated a high lexical coverage in this and similar balanced corpora. We were less sure about how our system would perform when dealing with narrow domain data, such as the METER Corpus, as the nature of such corpora ensures an abundance of technical terms and jargon. Our aim, then, was to test the impact of such features of data on the lexical coverage of our semantic lexicon and tagger.

As one might expect, the lexical coverage dropped when processing the Meter Corpus. However, the drop was only minimal (i.e. to 95.38%). Indeed, only 11,143 out of the total 241,311 tokens were found unmatched by our lexicon. After careful examination we identified two main reasons for this drop of lexical coverage:

1) The frequent use of domain-specific terminology in the METER Corpus, and
2) The frequent use of many of these unmatched terms, due to the homogeneous feature of the corpus.

Thus far, we have concentrated on synchronic factors. However, as previously highlighted, we were also interested in the diachronic factors that may affect the lexical coverage, even though we were aware that a number of factors would make a diachronic investigation difficult (i.e. differences in spelling practices, morphological inconsistencies, archaic/rare terminology)

As explained above, we drew our first historical test corpus from the Lancaster Newsbook Corpus (LNC), and tagged it using the same semantic lexicon that was used for modern English. The result obtained on the LNC was comparable to that on the METER Corpus, as shown in Table 2 below. To be precise, 3,418 out of the total 61,065 tokens were unmatched, resulting in a lexical coverage of 94.40%. The historical lexical variants cause similar reduction in the lexical coverage to that of domain-specific technical terms and jargon.

Test Corpus	Total Tokens	Unmatched Tokens	Lexical Coverage
Lancaster Newsbooks	61,065	3,418	94.40%
Gulliver's Travels	194,987	14,117	92.76%
Tristram Shandy	108,137	3,235	97.01%
Tom Jones	352,942	11,944	96.62%
Clarissa	887,276	40,988	95.38%
19th century fiction	5,000,000	135,661	97.29%.

Table 2: Lexical coverage on historical test corpora

We used the second test corpus (see *Gulliver's Travels*, *Tristram Shandy*, *Tom Jones*, *Clarissa* the 19th century fiction in Table 2) slightly differently than we had the first. As Archer *et al*. (2003) reported, we are developing an historical tagger by adding additional 'historical' lexicons to the existing semantic tagger (that is, a single lexicon dictionary and MWE dictionary, which contain items that are peculiar to earlier periods of English). In this part of the evaluation, we tagged the texts using both the modern lexicons and the newly developed historical lexicons.

In terms of the 18th century material, the range of lexical coverage showed considerable variation depending on the features of the various novels (i.e.. between 92.36% and 97.01% – see Table 2). For example, in the book, *Gulliver's Travels*, 14,117 out of the total 194,987 tokens were not matched, resulting in a lexical coverage of 92.76%. In *Tristram Shandy*, the lexicon coverage was slightly higher, that is, the system failed to identify 3,235 out of the total 108,137 tokens, producing a lexical coverage of 97.00%. We believe that the excessive use of nonce forms (Houyhnhnms, Glumdalcclitch, Blefuscu) may account for the higher error rate in *Gulliver's Travels.*

The expanded lexicon achieved even higher lexical coverage for the 19th century fiction, i.e. 97.29%. This means that only 135,661 out of the total 5 million tokens were unmatched by the USAS system. Higher lexical coverage for the 19th century fiction is most likely due to a more stable spelling system.

5. Discussion

As shown in the previous section, the Lancaster semantic lexicon has a wide lexical coverage for both

modern and historical English. Although the result is not conclusive due to the limitation of the scope of the test corpora, our evaluation shows that the Lancaster semantic lexicon of the USAS semantic tagger is capable of processing a wide range of corpus data. Nonetheless, the current Lancaster semantic lexicon has its limitations.

Firstly, although it obtained an extremely high lexical coverage of 99.39% on the BNC spoken data, when dealing with written texts, the lexical gap of about 2.4% or greater was persistent across the written test corpora. This unmatched part of the lexicon, although small, may include important key words that are critical for corpus analysis.

Another problem lies in the fact that our semantic lexicon is still not efficient in dealing with texts from specific domains. In fact, the METER corpus (our specific domain test corpus) is not drastically diverse from general English, as the texts are journalistic reports written for ordinary readers. Much more domain specific technical terms and jargon can be expected if we process more specific texts, such as collections of academic papers, business documents, etc.

The historical data presents an even tougher challenge. As our evaluation reveals, the lexical coverage of our semantic lexicon coverage fluctuates depending on the text. For example, the lexical coverage reached 97.29% on 19th century fiction whereas it dropped to 92.76% on *Gulliver's Travels*. And lexical coverage will probably worsen the further back we go (due to the use of now rare/archaic terminology and inconsistent spelling conventions). That said, we deliberately chose 18th century texts that we knew would prove problematic (because of the excessive use of nonce forms, the presence of morphological inconsistencies and the idiosyncratic style of the authors). Indeed, as Archer *et al*. (2003) reported, our results suggest that the USAS system (including its transition probability matrix) can adequately account for the grammatical features of Early Modern English, and lexical coverage can be improved by expanding the historical single item and MWE lexicons.

There is no easy answer to these problems. However, as relentless expansion of the lexicons may not always be practical, we are looking into other ways of collecting, structuring and applying lexical items to our system.

6. Conclusion

We have shown that a variety of factors can affect the lexical coverage of a lexicon, including genre, domain, date of publication, etc. We therefore contend that one needs to evaluate the potential of annotation systems and the coverage of their lexicons both diachronically and synchronically across genres, before making claims about their proficiency.

In terms of our own system, the Lancaster semantic lexicon, or the semantic tagging system, achieved a remarkably high coverage in modern general English language, the spoken genre in particular. The coverage degrades slightly when processing highly domain-specific or historical corpora, but the result is still encouraging. We will continue to expand our lexicon to improve the coverage, and make the system as 'future proof' as possible.

Acknowledgement

The work presented in this paper was carried out within the Benedict Project funded by the European Community under the 'Information Society Technologies' Programme (reference number: IST-2001-34237).

References

Archer, D., T. McEnery, P. Rayson, A. Hardie. (2003). Developing an automated semantic analysis system for Early Modern English. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16 (pp. 22 - 31). UCREL, Lancaster University.

Demetriou. George and Eric Atwell (2001). A domain-independent semantic tagger for the study of meaning associations in English text. In Proceedings of the 4th International Workshop on Computational Semantics (IWCS 4) (pp. 67-80). Tilburg, Netherlands.

Fellbaum, Christiane (ed.) (1998). WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press.

Gaizauskas, Robert, Jonathan Foster, Yorick Wilks, John Arundel, Paul Clough and Scott Piao (2001). The METER corpus: a corpus for analysing journalistic text reuse. In the Proceedings of Corpus Linguistics 2001 (pp: 214-223). Lancaster, UK.

Piao, Scott S. L., Paul Rayson, Dawn Archer, Andrew Wilson and Tony McEnery (2003). Extracting Multiword Expressions with a Semantic Tagger. In proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL 2003, 41st Annual Meeting of the Association for Computational Linguistics (pp. 49-56). Sapporo, Japan.

Rayson, P. and A. Wilson. (1996). The ACAMRIT semantic tagging system: progress report. In L. J. Evett and T. G. Rose (eds.) Language Engineering for Document Analysis and Recognition (LEDAR), AISB96 Workshop proceedings (pp. 13 – 20). Brighton, England. Faculty of Engineering and Computing, Nottingham Trent University, UK.

Vossen, P. (ed.) (1998). EuroWordNet: *A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht.