Measuring Semantic Coverage

Sergei Nirenburg, Kavi Mahesh and Stephen Beale
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003-0001
USA
sergei;mahesh;sb@crl.nmsu.edu

Abstract

The development of natural language processing systems is currently driven to a large extent by measures of knowledge-base size and coverage of individual phenomena relative to a corpus. While these measures have led to significant advances for knowledge-lean applications, they do not adequately motivate progress in computational semantics leading to the development of large-scale, general purpose NLP systems. In this article, we argue that depth of semantic representation is essential for covering a broad range of phenomena in the computational treatment of language and propose depth as an important additional dimension for measuring the semantic coverage of NLP systems. We propose an operationalization of this measure and show how to characterize an NLP system along the dimensions of size, corpus coverage, and depth. The proposed framework is illustrated using several prominent NLP systems. We hope the preliminary proposals made in this article will lead to prolonged debates in the field and will continue to be refined.

1 Measures of Size versus Measures of Depth

Evaluation of current and potential performance of an NLP system or method is of crucial importance to researchers, developers and users. Current performance of systems is directly measured using a variety of tests and techniques. Often, as in the case of machine translation or information extraction, an entire "industry" of evaluation gets developed (see, for example, ARPA MT Evaluation; MUC-4). Measuring the performance of an NLP method, approach or technique (and through it the promise of a system based on it) is more difficult, as judgments must be made about "blame assignment" and the impact of improving a variety of system components on the overall future performance. One of the widely accepted measures of potential performance improvement is the feasibility of scaling up the static knowledge sources of an NLP system - its grammars, lexicons, world knowledge bases and other sets of language descriptions (the reasoning being that the larger the system's grammars and lexicons, the greater percentage of input they would be able to match and, therefore, the better the performance of the system[1]). As a result, a system would be considered very promising if its knowledge sources could be significantly scaled up at a reasonable expense. Naturally, the expense is lowest if acquisition is performed automatically. This consideration and the recent resurgence of corpus-based methods heighten the interest in the automation of knowledge acquisition. However, we believe that such acquisition should not be judged solely by the utility of acquired knowledge for a particular application.

A preliminary to the scalability estimates is a judgment of the current coverage of a system's static knowledge sources. Unfortunately, judgments based purely on size are often misleading. While they may be sufficiently straightforward for less knowledge-intensive methods used in such applications as information extraction and retrieval, part of speech tagging, bilingual corpus alignment, and so on, the same is not true about more rule- and knowledge-based methods (such as syntactic parsers, semantic analyzers, semantic lexicons, ontological world models, etc.). It is widely accepted, for instance, that judgments of the coverage of a syntactic grammar in terms of the number of rules are flawed. It is somewhat less self-evident, however, that the number of lexicon entries or ontology concepts is not an adequate measure of the quality or coverage of NLP

[1] Incidentally, this consideration contributes to evaluation of current performance as well. In the absence of actual evaluation results, it is customary to claim the utility of the system by simply mentioning the size of its knowledge sources (e.g., "over 550 grammar rules, over 50,000 concepts in the ontology and over 100,000 word senses in the dictionary").

systems. An adequate measure of these must examine not only size and its scalability, but also depth of knowledge along with its scalability. In addition, these size and depth measures cannot be generalized over the whole system, but must be directly associated with individual areas that cover the breadth of NLP problems (i.e. morphology, word-sense ambiguity, semantic dependency, coreference, discourse, semantic inference, etc.). And finally, the most helpful measurements will not judge the system solely as it stands, but must in some way reflect the ultimate potential of the system, along with a quantification of how far additional work aimed at size and depth will bring about advancement toward that potential.

In this article, we attempt to formulate measures of coverage important to the development and evaluation of semantic systems. We proceed from the assumption that coverage is a function of not only the number of elements in (i.e., **size** of) a static knowledge source but also of the amount of information (i.e., **depth**) and the types of information (i.e., **breadth**) contained in each such element. Static size is often emphasized in evaluations with no attention paid to the often very insignificant amount of information associated with each of the many "labels" or primitive symbols. We suggest a starting framework for measuring size together with other significant dimensions of semantic coverage. In particular, the evaluation measures we propose reflect the necessary contribution of the depth and breadth of semantic descriptions. Depth and breadth of semantic description are essential for progress in computational semantics and, ultimately, for building large-scale, general purpose NLP systems. Of course, for a number of applications a very limited semantic analysis (e.g., in terms of, say, a dozen separate features) may be adequate for sufficiently high performance. However, in the long run, progress towards the ultimate goal of NLP is not possible without depth and breadth in semantic description and analysis.

There is a well-known belief that it is not appropriate to measure success of NLP using field-internal criteria. Its adherents maintain that NLP should be evaluated exclusively through evaluating its applications: information retrieval, machine translation, robotic planning, human-computer interaction, etc. (see, for example, the Proc. of the Active NLP Workshop; ARPA MT Evaluation). This may be true for NLP users, but developers must have internal measures of success. This is because it is very difficult to assign blame for the success or failure of an application on specific components of an NLP system. For example, in reporting on the MUC-3 evaluation efforts, Lehnert and Sundheim (1991) write:

> A wide range of language processing strategies was employed by the top-scoring systems, indicating that many natural language-processing techniques provide a viable foundation for sophisticated text analysis. Further evaluation is needed to produce a more detailed assessment of the relative merits of specific technologies and establish true performance limits for automated information extraction. [emphasis added.]

Thus, evaluating the information extraction application did not provide constructive criticism on particular NLP techniques to enable advances in the state of the art. Also, evaluating an application does not directly contribute to progress in NLP as such. This is in part because a majority of current and exploratory NLP systems are not complete enough to fit an application but rather are devoted to one or more of a variety of components of a comprehensive NLP system (static – e.g., lexicons, grammars, etc.; or dynamic – e.g., an algorithm for treating metonymy in English).

1.1 Current Measures of Coverage

Success in NLP (including semantic analysis and related areas) is currently measured by the following criteria:

- Size of static knowledge sources: A mere number indicating the size of a knowledge source does not tell us much about the coverage of the system, let alone its semantic capabilities. For example, most machine readable dictionaries (MRD) are larger than computational lexicons but they are not usable for computational semantics.
- Coverage of corpus, either blanket coverage ("56% of sentences were translated correctly") or resolution of a certain phenomenon ("78% of anaphors were determined correctly"). These measures are often misleading by themselves since what may be covered are just one or two highly specific phenomena such as recognizing place or product names (i.e., limited breadth). NLP is not yet at a stage where "covering a corpus" can mean "analyzing all elements of meanings of texts in the corpus." It may be noted that "correctly" is a problematic term since people often have difficulty judging what is "correct" (Will, 1993). Moreover, correctness is orthogonal to the entire discussion here since we would like to increase semantic coverage along various dimensions while maintaining an acceptable degree of correctness. On the same lines, processing efficiency (often specified in terms such as "A sentence of length 9 takes 750 milliseconds to process") is also more or less orthogonal to the dimensions we propose for measuring semantic coverage. Increasing semantic coverage would be futile if

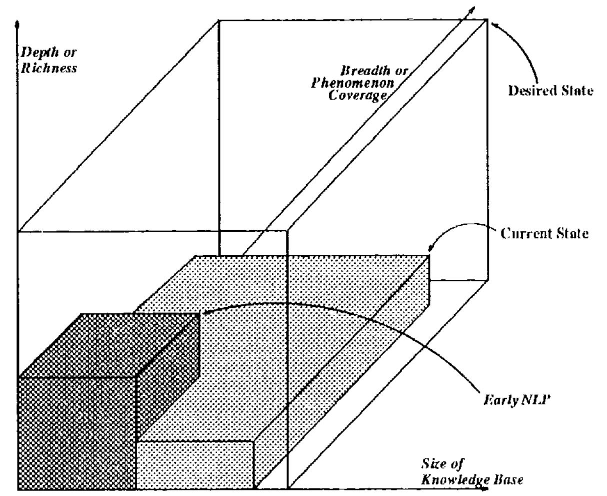

Figure 1: Dimensions of Semantic Coverage: Current and Desired Directions

processing became exponentially expensive as a result.

Figure 1 shows the dimensions of size and breadth (or phenomenon coverage) along the horizontal plane. Depth (or richness) of a semantic system is shown on the vertical axis. We believe that recent progress in NLP with its emphasis on corpus linguistics and statistical methods has resulted in a significant spread along the horizontal plane but little been done to grow the field in the vertical dimension. Figure 1 also shows the desired state of computational semantics advanced along each of the three dimensions shown. If

We proceed from the assumption that high-quality NLP systems require optimum coverage on all three scales, then apparently different roads can be taken to that target. The spectrum of choices ranges from developing all three dimensions more or less simultaneously to taking care of them in turn. As is often the case in long-term high-risk enterprises, many researchers opt to start out with acquisition work which promises short-term gains on one of the coverage dimensions, with little thought about further steps. Often the reason they cite can be summarized by the phrase "Science is the art of the possible." This position is quite defensible — if no claims are made about broad semantic coverage. Indeed, it is quite legitimate to study a particular language phenomenon exclusively or to cover large chunks of the lexis of a language in a shallow manner. However, for practical gains in large-scale computational-semantic applications one needs to achieve results on each of the three dimensions of coverage.

1.2 Desiderata for Large-Scale Computational Semantics

Once the initial knowledge acquisition campaign for a particular application has been concluded, the following crucial scalability issues[2] must be addressed, if any understanding of the longer-term significance of the research is sought:

- domain independence: scalability to new domains; general-purpose NLP
- language independence: scalability across languages
- phenomenon coverage: scalability to new phenomena; going beyond core semantic analysis; ease of integrating component processes and resources.
- application-independence: scalability to new applications; toolkit of NLP techniques applicable to any task.

We believe that coverage in terms of the depth and breadth of the knowledge given to an NLP system is mandatory for attaining the above goals in the long run. Such coverage is best estimated not in terms of raw sizes of lexicons or world models but rather through the availability in them of information necessary for the treatment of a variety of phenomena in natural language — issues related to semantic dependency building, lexical disambiguation, semantic constraint tracking and relaxation (for the cases of unexpected input, including non-literal language as well as treatment of unknown lexis), reference, pragmatic impact and discourse structure. The resolution of these issues is at the core of post-syntactic text processing. We believe that one can treat the above phenomena only by acquiring a broad range of relevant knowledge elements for the system. One useful measure for sufficiency of information would be an analysis of kinds of knowledge necessary to generate a text (or dialog) meaning representation. For applications in which more procedural computational semantics is preferable, a corresponding measure of sufficiency should be developed.

There exist other, broader desiderata which are applicable to any AI system. They include concerns about system robustness, correctness, and efficiency which are orthogonal to the above issues. Equally important but more broadly applicable are considerations of economy and ease of acquisition of knowledge sources — for example, reducing the size of knowledge bases and sharing knowledge across applications.

[2] At present, scalability is considered in the field almost exclusively as propagation of the number of entries in the NLP knowledge bases, not the quantity and quality of information inside each such entry.

2 How to Reason about Depth, Breadth and Size

A useful measure of semantic coverage must involve measurement along each of the three dimensions with respect to correctness (or success rate) and efficiency (or speed). In this first attempt at a qualitative metric, we list questions relevant for assigning qualitative ("tendency") scores to an NLP system to measure its semantic coverage. Our experience over the years has led us to the following sets of criteria for measuring semantic coverage. However, we understand that the following are not complete or unique; they are representative of the types of issues that are relevant to measuring semantic coverage.

2.1 Lexical Coverage

- To what extent do entries share semantic primitives (or concepts) to represent word meanings? What is the relation between the number of semantic primitives defined and the number of word senses covered?
- What is the size of the semantic zones of the entry? How many semantic features are covered?
- How many word senses from standard human-oriented dictionaries are covered in the NLP-oriented lexicon entry?
- What types of information are included?
  - selectional restrictions
  - constraint relaxation information
  - syntax-semantics linking
  - collocations
  - procedural attachments for contextual processing
  - stylistic parameters
  - aspectual, temporal, modal and attitudinal meanings
  - other idiosyncratic information about the word
- and, finally, the total number of entries in the lexicon.

2.2 Ontological Coverage

The total number of primitive labels in a world model is not a useful measure of the semantic coverage of a system. At least the following considerations must be factored in:

- The number of properties and links defined for an individual concept
- Number of types of non-taxonomic relationships among concepts
- Average number of links per concept: "connectivity"
- Types of knowledge included: defaults, selectional constraints, complex events, etc.
- Ratio of number of entries in a lexicon to number of concepts in the ontology
- and, finally, total number of concepts in the ontology.

2.3 Measuring Breadth of Meaning Representations

Apart from lexical and ontological coverage, the depth and breadth of the meaning representations constructed by a system are good indicators of the overall semantic coverage of the system. The number of different types of meaning elements included from the following set provides a reasonable measure of coverage:

- Argument structure only
- Template filling only
- Events and participants
- Thematic role assignments
- Time and temporal relations
- Aspect
- Properties: attributes of events and objects; relations between events and objects.
- Reference and coreference
- Attitude, modality, stylistics
- Quantitative, comparative, and other mathematical relations
- Textual relations and other discourse relations
- Multiple ambiguous interpretations
- Propositional and story/dialog structure

3 Measuring Semantic Coverage: Examples

Figure 2 shows the approximate position of several well-known approaches and systems (including a possible Cyc-based system) in the 3-dimensional space of semantic coverage. We have chosen representative systems from the different approaches for lack of precise terms to name the approaches.

How do the approaches illustrated in Figure 2 rate with respect to the metrics suggested above? When estimating their profiles, we thought either about some representative systems belonging to an approach or thought of the properties of a prototypical system in a particular paradigm if no examples presented themselves readily. In the interests of space, we consider the above criteria for measuring semantic coverage but only provide brief summaries of how each system or approach is located along the dimensions of depth, breadth and size.

The schema-based reasoner, Boris (Lehnert et al, 1983) was used as a prototype system for the

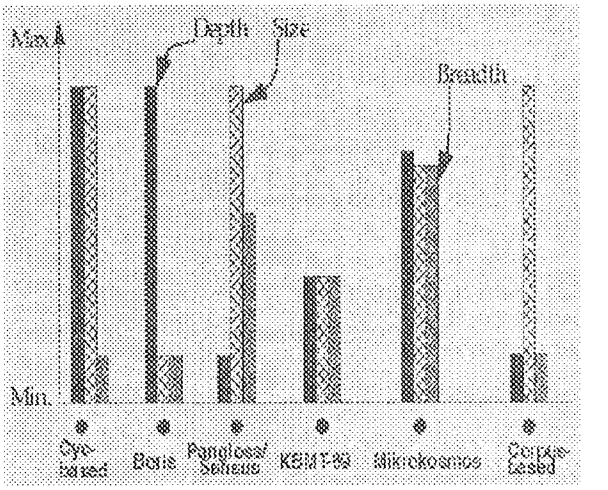

Figure 2: Dimensions of Semantic Coverage: Current and Desired Directions

domain- and task-dependent, AI-style, schema-based NLP system. It may be considered an extreme example of a system with deep, rich knowledge of its, rather narrow, world in which covering language phenomena is needed only inasmuch as it supports general reasoning. Boris was able to process a very small number of texts sufficiently for its goals. The coverage of phenomena was strictly utilitarian (which, we believe, is quite appropriate). It was not demonstrated that Boris can be scaled up to cover a significant part of the English lexicon.

As an example of an early knowledge-based MT system (that is, unlike the above, a system whose goals were mainly computational-linguistic) we chose the KBMT-89 system (Goodman and Nirenburg, 1991). It covered its small corpus relatively completely and described the necessary phenomena relatively fully. It was a primary goal of this line of research to begin meeting the above criteria for semantic coverage.

A putative NLP system based on the Cyc project has been selected as a prototype for systems not devised for a particular application. The Cyc large-scale knowledge base has significant amounts of deep knowledge. However, it is not clear whether the knowledge is applicable in a straightforward manner to deal with a range of linguistic phenomena. The big question for this kind of system is whether it is, in fact, possible, to acquire knowledge without a reference to an intended application.

A purely corpus-based, statistical approach to NLP, on the other hand, has an extremely narrow range of knowledge, but may have a large size. For example, such a system may have a large lexicon with only word frequency and collocation information in each entry. Although statistical methods have been shown to work on some problems and applications, they are typically applied to one or two phenomena at a time. It is not clear that statistical information acquired for one problem (such as sense disambiguation) is of use in handling other problems (such as processing non-literal expressions).

Mixed-strategy NLP systems are epitomized by Pangloss (1994), a multi-engine translation system in which semantic processing is only one of the possible translation engines. The semantics engine of this system is equipped with a large-size ontology of over 50,000 entries (Knight and Luk, 1994) which is used essentially as an anchor for mapping lexical units from the source to the target language. As shown in Figure 2, Pangloss has a large size and covers a good range of phenomena as well. However, there is little information (only taxonomic and partonomic relationships) in each concept in its Sensus ontology. The limited depth constrains the ultimate potential of the system as a semantic and pragmatic processor. For example, there is no information in its knowledge sources to make judgements about constraint relaxation to process non-literal expressions such as metonymies and metaphors.

The Mikrokosmos system (e.g., Onyshkevych and Nirenburg, 1994), has attempted to cover each dimension equally well. Its knowledge bases and text meaning representations are rather deep and of nontrivial sizes. It has been designed from the start to deal with a comprehensive range of semantic phenomena including the linking of syntax and semantics, core semantic analysis, sense disambiguation, processing non-literal expressions, and so on, although not all of them have yet been implemented.

From the above examples, it is clear that having good coverage along one or two of the three dimensions is not good enough for meeting the long-term goals of NLP. Poor coverage of language phenomena (i.e., poor breadth) indicates that the acquired knowledge, even when it is deep and large in size, may not be applicable to other phenomena and may not transfer to other applications. Poor depth suggests that knowledge and processing techniques are either application- or language-specific and limits the ultimate potential of the system in solving semantic problems. Depth and breadth are of course of little use if the system cannot be scaled up to a significant size. Moreover, as already noted, coverage in depth, breadth, and size must all be achieved in conjunction with maintaining good measures of correctness, efficiency, and robustness.

4 Discussion and Conclusions

An oft-quoted objection to having deep semantic coverage is the difficulty in scaling up such a system along the dimension of size. This is a valid concern. However, the situation can be ameliorated to a large extent by developing a methodology (see, e.g., Mahesh and Nirenburg, 1995) for constraining knowledge acquisition to minimally meet semantic processing needs. Such concentration of effort will allow knowledge acquirers to have spend a fraction of the effort that must go into building a general machine-tractable encyclopedia of knowledge and yet to attain significant coverage of language phenomena. Significant scale-up can be accomplished under such a constraint without jeopardizing the high values on the depth and breadth scales.

Size is important in NLP. But size alone is not a sufficient metric for evaluating semantic coverage. Focusing on size to the exclusion of other criteria has biased the field away from semantic solutions to NLP problems. We have made a first step in formulating a more appropriate and complete set of measures of semantic coverage. Depth and breadth of knowledge necessary to cover a wide range language phenomena are at least as important to NLP as size. The discussion of peculiarities of the various approaches should be expanded in at least two directions – greater detail of description and analysis of the relative difficulty of reaching the set goal of attaining an optimum value on each of the three measurement scales. We hope that this paper will elicit interest in continued discussion of the issues of coverage measurement, which, in turn, will lead to better – quantitative as well as qualitative – measures, including a methodology for comparing lexicons and ontologies.

Acknowledgments
Many thanks to Yorick Wilks for his constructive criticism.

References

Active NLP Workshop: Working Notes from the AAAI Spring Symposium "Active NLP: Natural Language Understanding in Integrated Systems" March 21-23, 1994, Stanford University, California (Also available as a Technical Report from the American Association for Artificial Intelligence).

ARPA MT Evaluation: Report of the Advanced Research Projects Agency, Machine Translation Program System Evaluation, May-August 1993.

Goodman, K. and S. Nirenburg (eds.) (1991). *The KBMT Project: A Case Study in Knowledge-Based Machine Translation.* San Mateo, CA: Morgan Kaufmann.

Knight, K. and Luk, S. K. (1994). Building a Large-Scale Knowledge Base for Machine Translation. In Proc. Twelfth National Conf. on Artificial Intelligence, (AAAI-94).

Lenat, D. B. and Guha, R. V. (1990). *Building Large Knowledge-Based Systems.* Reading, MA: Addison-Wesley.

Lehnert, W. G., Dyer, M. G., Johnson, P. N., Yang, C. J., and Harley, S. (1983). BORIS - An Experiment in In-Depth Understanding of Narratives. *Artificial Intelligence*, 20(1):15-62.

Lehnert, W. G. and Sundheim, B. (1991). A performance evaluation of text-analysis technologies. *AI Magazine*, 12(3):81-94.

Mahesh, K. and Nirenburg, S. (1995). A situated ontology for practical NLP. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada, August 1995.

MUC-4: Proc. Fourth Message Understanding Conference (MUC-4), June 1992. Defense Advanced Research Projects Agency.Morgan Kaufmann Publishers.

Onyshkevych, B. and Nirenburg, S. (1994). The lexicon in the scheme of KBMT things. Technical Report MCCS-94-277, Computing Research Laboratory, New Mexico State University. Also to appear in Machine Translation.

Pangloss. (1994). The PANGLOSS Mark III Machine Translation System. A Joint Technical Report by NMSU CRL, USC ISI and CMU CMT, Jan. 1994.

Will, C. A. (1993). Comparing human and machine performance for natural language information extraction: Results from the Tipster evaluation. Proc. Tipster Text Program, ARPA, Morgan Kaufmann Publishers.