# Evaluating Complex Semantic Artifacts

**Christopher R Walker, Hannah Copperman**
Powerset, a Microsoft company
475 Brannan St, Suite 330, San Francisco, CA, 94107, USA
{chriwalk,hannahc}@microsoft.com

**Abstract**

Evaluating complex Natural Language Processing (NLP) systems can prove extremely difficult. In many cases, the best one can do is to evaluate these systems indirectly, by looking at the impact they have on the performance of the downstream use case. For complex end-to-end systems, these metrics are not always enlightening, especially from the perspective of NLP failure analysis, as component interaction can obscure issues specific to the NLP technology. We present an evaluation program for complex NLP systems designed to produce meaningful aggregate accuracy metrics with sufficient granularity to support active development by NLP specialists. Our goals were threefold: to produce reliable metrics, to produce useful metrics and to produce actionable data. Our use case is a graph-based Wikipedia search index. Since the evaluation of a complex graph structure is beyond the conceptual grasp of a single human judge, the problem needs to be broken down. Slices of complex data reflective of coherent Decision Points provide a good framework for evaluation using human judges (Medero et al., 2006). For NL semantics, there really is no substitute. Leveraging Decision Points allows complex semantic artifacts to be tracked with judge-driven evaluations that are accurate, timely and actionable.

## 1. Background

Evaluating complex Natural Language Processing (NLP) systems can prove extremely difficult. In many cases, the best one can do is to evaluate these systems indirectly, by looking at the impact they have on the performance of the downstream use case. For complex end-to-end systems, these metrics are not always enlightening, especially from the perspective of NLP failure analysis, as component interaction can obscure issues specific to the NLP technology. We present an evaluation program for complex NLP systems designed to produce meaningful aggregate accuracy metrics with sufficient granularity to support active development by NLP specialists.

Our use case is a graph-based Wikipedia search index. The document index is generated by a multi-step NLP system. The search algorithm uses a similar NLP system on the query side, and then searches for a graph match in the index. Both query- and index-side NLP systems have three linguistic components: a high-precision Named Entity Tagger; an LFG-based parser (Kaplan et al., 1996; PARC), which produces the C- and F-Structures for the most likely parse; and a Semantic Transfer system (Crouch et al., 2006), which generalizes the parse-graphs further to account for well-understood lexical relations, structural paraphrases and various other semantic features. The output of the entire NLP system is a batch of documents in an XML variant that we refer to as SemXML.

When applying NLP technology in a feature-driven environment, such as consumer search, the metrics used to drive overall progress are often not well-suited to the problems of NLP in particular. For example, any metric designed to capture the overall relevance of the search engine results, such as NDCG (Jarvelin et al., 2002), will be only marginally actionable for NLP developers. There are simply too many variables, many of which have nothing to do with linguistic processing or with semantic matching.

Our situation is complicated by the fact that SemXML is an artifact of complex interactions between NLP components, where the performance of a single NLP component will not paint a complete picture of the overall performance of the system. Only a careful inspection of the SemXML artifact itself can provide an adequate framework for system evaluation.

We present here a complete program for SemXML Evaluation. In the pursuit of an evaluation framework for complex system-generated semantic graphs, our goals were threefold: to produce *reliable metrics*, to produce *useful metrics* and to produce *actionable data*.

## 2. Decision Points

Since the evaluation of a complex graph structure is beyond the conceptual grasp of a single human judge, the problem needs to be broken down. Slices of complex data reflective of coherent *Decision Points* provide a good framework for evaluation using human judges (Medero et al., 2006).

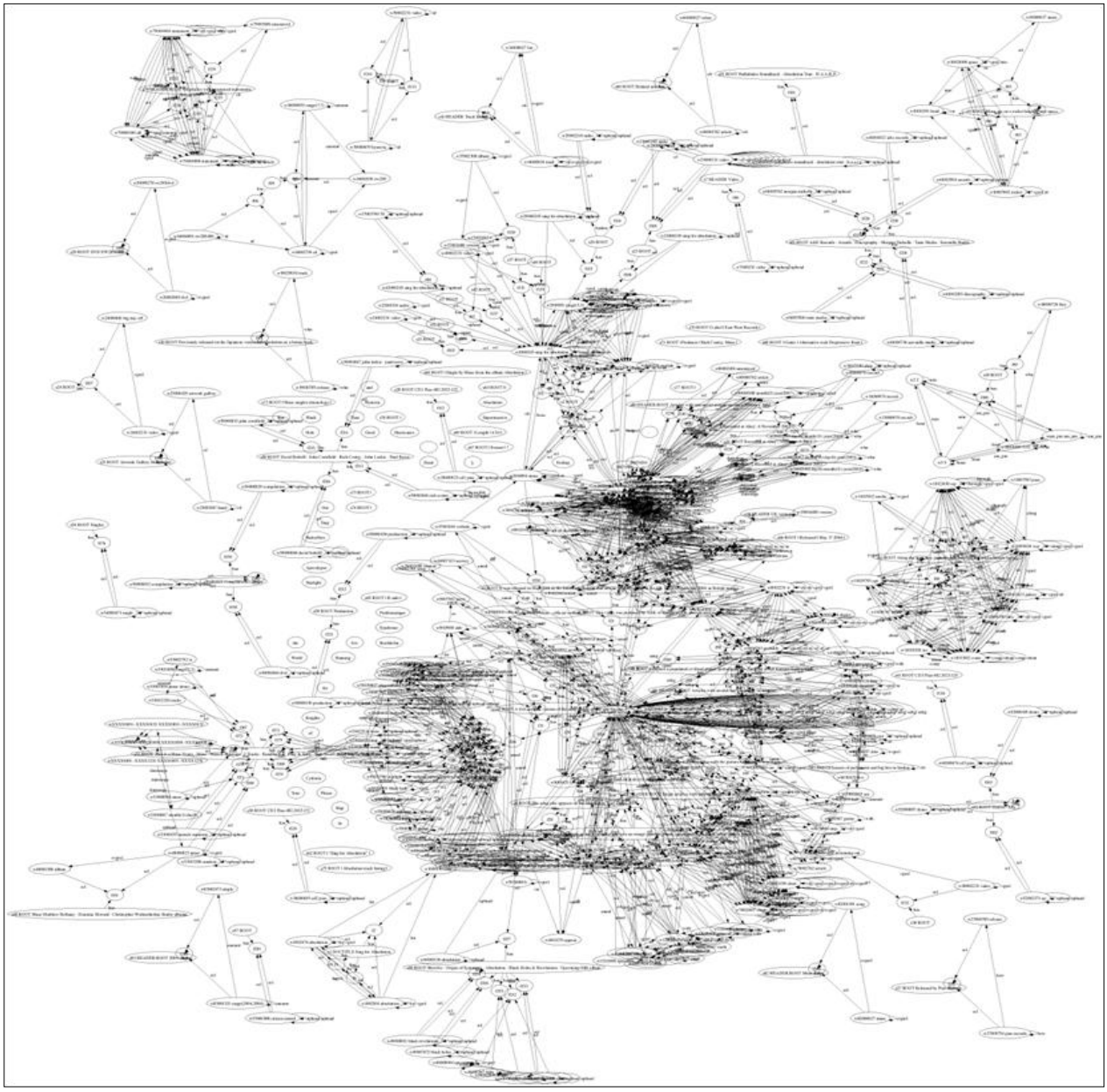


Figure 1: SemXML Graph Representation for a Wikipedia Document

We take classes of nodes and edges as the appropriate objects of evaluation for a single human judgment task. Although the terminology has been simplified a bit from the actual index model, we will use the term *semword* to refer to a unique node in the graph. Edge types can be any number of grammatical and semantic relations. A small cluster of edges rooted at a single node is referred to as a *fact*—its root node is called the *relation*.

A simplified fact for the sentence *John loves Mary* might be represented in the SemXML graph as follows:

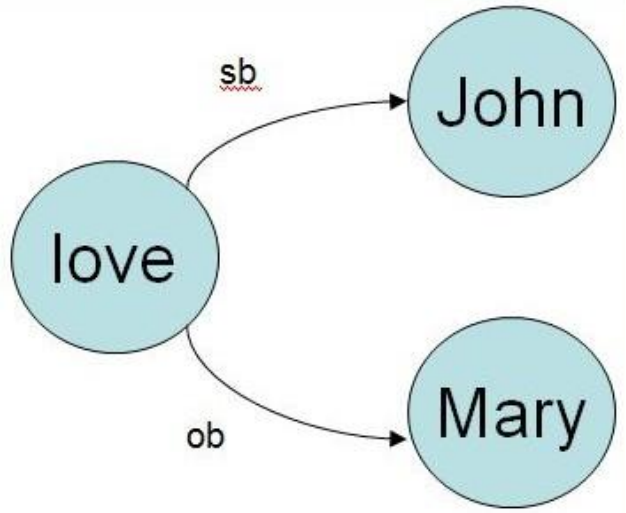



Figure 2: Simplified Fact Graph

Here, there are three semword nodes (*John*, *love* and *Mary*). The special semword in this fact, the relation, is *love*. Given this graph, we would want to evaluate two Decision Points: (1) is the *sb* edge correct? and (2) is the *ob* edge correct?

For the purposes of our evaluation, there are four classes of Decision Points (DPs):

1. **Role Assignments:** This class of DPs focuses on an edge connecting two semwords in the same fact. The edge is labeled with the role being judged and one of the nodes is the fact's special relation semword. Here the correctness of asserting the labeled edge is assessed. Among the types of linguistic judgments involved in these assignments are: grammatical roles such as subject (*sb*) and direct object (*ob*); and asserted identity, both copular and appositional (*eid* and *id*, respectively).

2. **Name Tags:** This class of DPs focuses on the features of a single semword—specifically, the correctness of its nametag hypothesis (i.e. span and type).

3. **Re-verbal Inferences:** This class of DPs focuses on a fact's relation semword. These cases are judged correct only in case the semword (whose original POS is *noun*) should also be interpreted with the associated verbal argument structure. For example, we would want to know the correct argument structure for the verb *destroy* if we encountered the expression *the destruction of the city by Godzilla*. Note that the judgment here is whether or not we should even look for the verbal arguments at all (their correctness is evaluated with all the other role assignments).

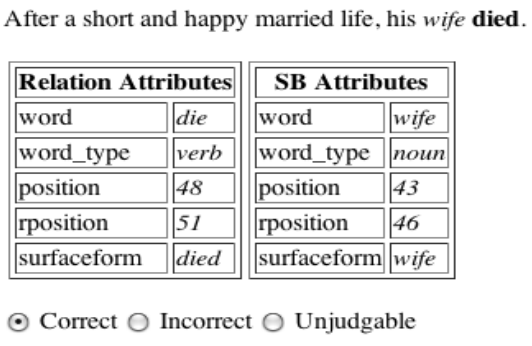
4. **Co-reference Links:** This class of DPs focuses on a pair of semwords. Since the two evaluated semwords are often in two different sentences, the scope of the evaluated sub-graph is actually larger than a single fact. The extra-factual link is provided by a unique identifier, called a *skolem*. For some semword types, two semwords with the same skolem are taken to co-refer. We treat this judgment as analogous to other edge judgments.

## 3. Evaluation Procedure

The evaluation procedure for SemXML must be complete enough that the entire range of possible errors are detected, but it also must be done quickly enough that the evaluation will still be useful and relevant once completed.

For each index evaluation, we take two data samples: one stable set (i.e., the same set of documents for each evaluation), and one random set. Each set contains approximately 1000 documents; each document is a single parsed and semantically processed Wikipedia article. A set this size is typically enough to produce 1000+ instances for even low-yield DPs. For DPs that yield more than one instance per document, we take the first 1000 we encounter.

Each example is presented to a judge with the relevant semwords emphasized (Copperman et al., 2010). A table under the sentence contains other linguistic information, such as stem, POS, etc. Judges are presented with three choices: *Correct*, *Incorrect* or *Unjudgable*:


After a short and happy married life, his *wife* **died**.

| Relation Attributes | | SB Attributes | |
|---|---|---|---|
| word | *die* | word | *wife* |
| word_type | *verb* | word_type | *noun* |
| position | *48* | position | *43* |
| rposition | *51* | rposition | *46* |
| surfaceform | *died* | surfaceform | *wife* |

◉ Correct ○ Incorrect ○ Unjudgable

Figure 3: Sample Judgment

Each batch of DP examples is evaluated by two trained judges, guided by detailed specifications. Judges are encouraged to note troublesome cases, since the specifications are always open to more revision.

Our current evaluation includes 12 DPs. With two separate sample sets of at least 1000 sentences each, 24,000 examples must be evaluated per index. Since every DP evaluation must be done by two judges, the total number of judgments is at least 48,000. Judges average roughly 200 judgments per hour, bringing the total volume of judge time required to a total of 240+ hours per evaluation. To meet the two-week target, an annotation team with 125 or more hours weekly commitment is required. Typically, ten hours per week is an upper bound on the reliable contribution of a single judge, so a team of at least 12 part-time judges is necessary.

## 4. SemXML Evaluation Metrics

Our procedure measures precision. Given the difficulty of establishing recall relative to a shifting semantic representation and constantly updated source material, we instead report *yield* as a proxy.

1. **Yield (Y):** The rate at which the phenomenon in question occurs, relative to the number of sentences or documents examined.
2. **Precision (P):** The rate at which the assertion of some indexed phenomenon is a correct assertion.

We define two DP-aggregating metrics:

1. **Mean Precision:** The average precision for the DPs evaluated.
2. **Y-Weighted Precision:** This metric de-emphasizes low-yield DPs by weighting the average precision for each DP with its yield:

$$\frac{\sum_{d \in DP} Y_d \times P_d}{\sum_{d \in DP} Y_d}$$

Each metric is applied to two sets of DPs:

1. **Core Roles:** A small set of important DPs evaluated since the inception of the SemXML Evaluation program.
2. **Major DPs:** All the major DPs evaluated as of the relevant index.

Core Roles include the relations most likely to be reflective of overall system performance: Identity (EID & ID), Subject (SB) and Object (OB). The use of these roles is partly historical, partly pragmatic. By choosing a small set of representative DPs, we are able to get a view of the general trends in aggregate SemXML precision. These four roles are particularly representative of the overall success of an NLP system in correctly identifying a wide range of complex predicate-argument relations.

The set of major DPs is extended for each evaluation cycle. As of the most recent evaluation, our set included 12 major DPs, ranging from the 4 Core Roles to more semantic relations such as Co-reference and Temporal Modification.

## 5. Metric Reliability

One aspect of our approach that has gone unmentioned is the absence of a "gold standard." For a number of reasons, such as the fluidity of Wikipedia source documents and the dynamic complexity of the underlying representation, such a corpus is not available. Nor is it possible to create a corpus with adequate coverage to measure Recall and of sufficient fuzziness to capture multiple possible "correct answers" (as is often necessary to accommodate different conventions for multi-token elements).

In the absence of well-defined gold standards, we cannot use the conventional approach of testing judges against existing data to confirm their task-readiness. Instead, we must treat judgment task stability and annotator preparedness as two sides of the same problem. To determine whether we can trust our precision metrics, we track three trends:

1. Inter-Annotator Agreement (IAA)
2. Precision variation across annotators
3. Precision variation across samples

We treat as reliable only those precision scores with at least 85% IAA. As of the most recent evaluation cycle, the mean IAA is 87.55% for core roles and 87.24% for all major DPs.

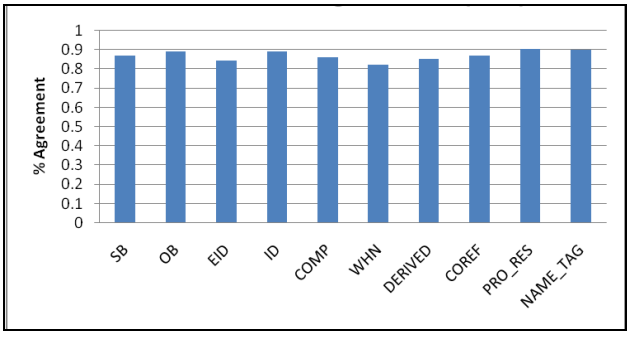



Figure 4: Summary of Inter-Annotator Agreement for Major Decision Points

Additionally, all but one Decision Point meet our standards. The DP in need of IAA improvements (*WHN*) is undergoing significant specification review in preparation for the next evaluation cycle. Focused spec development has driven IAA improvement for other roles. Early work on the judgment specifications for the core role *ID* improved IAA from 69.1% to 90.7% in a single evaluation cycle.

We also use other measures of task stability to drive spec development. Although some degree of variation across samples is to be expected, two independent samples should produce roughly the same results for the same <DP,index> pair. For well-defined DPs with acceptable IAA, this is what was seen:

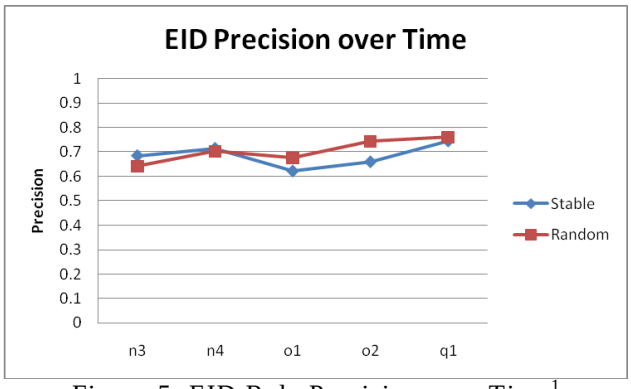



Figure 5: EID Role Precision over Time[1]

Interestingly, even DPs with less than 85% IAA produce results that are consistent enough to reflect trends in performance:

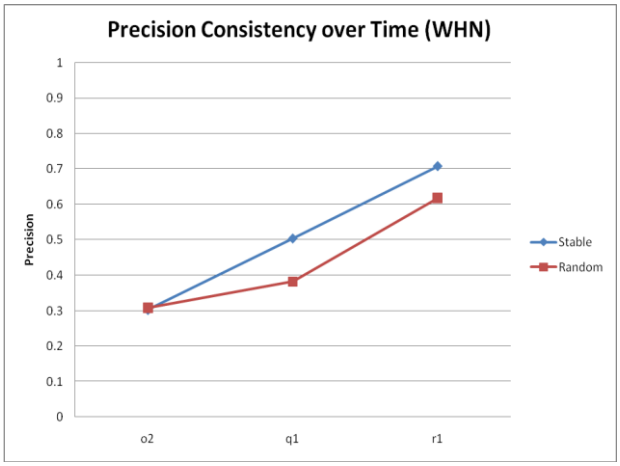



Figure 6: WHN Precision over Time

## 6. Metric Usefulness

Probably the most direct assessment of our metrics is their timeliness. Although developers often prefer a more immediately available metric, a two-week scale is more than adequate to the demands of longer-term development.

A more important, but less directly accessible, measure of utility is the presence of actual signal in the metrics. Once the evaluation tasks had stabilized, we were able to get clear signal from known improvements. For example, in the following assessment of Y-weighted Precision over time, we see two clear spikes: the first a result of work on Wikipedia text extraction; the second a result of focused development on two specific DPs:

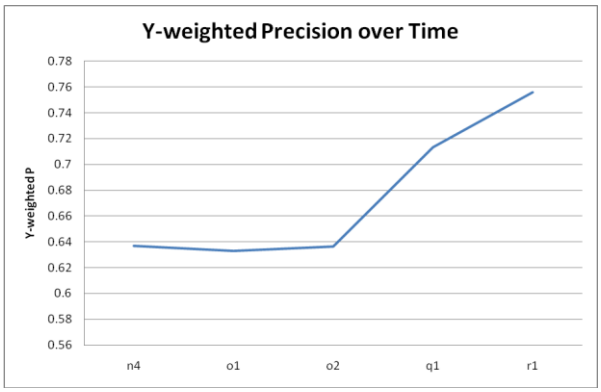



Figure 7: Y-Weighted Precision over Time

And these spikes correspond to more anecdotal reports of improvements in "system quality." It appears that precision can be improved when a metric is available. And when specific problem areas are targeted, improvements are reflected in the signal. Unexpected signal often proves to be the result of unforeseen interactions between system components.

Further analysis gives developers clues about what to fix. Analysis of the DP *WHN* resulted in a 20% absolute increase in precision for that DP. Similar results were seen for the DP *ID*. In both cases, partitioning judged data on the basis of its correctness provided a working DevTest set for rapid development and testing. More qualitative applications of partitioned data have also led to minor successes.

Finally, we provide a richly textured analysis of error distributions for each Decision Point, based on a number of sentence and document features (both linguistic and non-linguistic). For example, our precision for the recognition of clausal subject is distributed as follows:

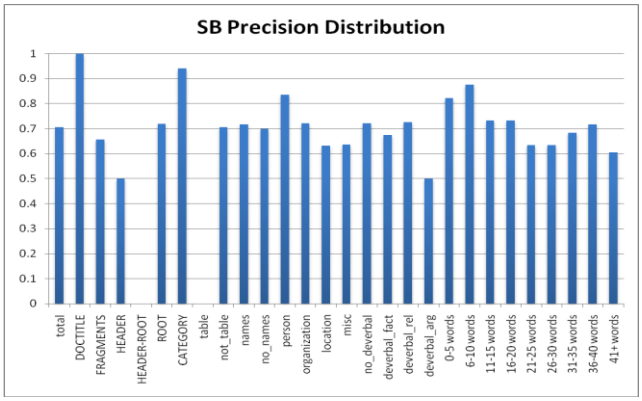



Figure 8: Sample Summary of Precision Distribution

Developers can use these distribution statistics to identify the precise classes of sentences that need additional analysis. Some of the classes are motivated by traditional linguistic features, such as the presence of proper names.

[1] Here "over time" is used to mean "from index build to index build". Each build can be roughly equated with two months elapsed calendar time.

But most are simply "computational" classes, such as the length of the relevant sentences.

In addition to providing developers with high-level summaries of DP error distributions, these analyses also allow for the partitioning of the judged examples themselves. Given such a robust collection of judged linguistic data organized across multiple dimensions of analysis, it is extremely easy for a developer to drill down into the corpora for additional failure analysis.

## 7. Conclusion

When it comes to NL semantics, there is no substitute for human judges. Leveraging Decision Points allows complex semantic artifacts to be tracked with judge-driven evaluations that are accurate, timely and actionable.

## 8. References


Copperman, H. and Walker, C. R. (2010). Fred's Reusable Evaluation Device: Providing Support for Quick and Reliable Linguistic Annotation. In *LREC 2010: Seventh International Conference on Language Resources and Evaluation.*

Crouch, D. and King, T.H. (2006). Semantics via F-Structure Rewriting. In *Proceedings of LFG06, CSLI On-line publications*, pp. 145-165.

Jarvelin, K. and Kekalainen, J. (2002). Cumulated gain-based evaluation of IR techniques. In *ACM Transactions on Information Systems* 20(4), 422–446.

Kaplan, R. M. and Bresnan, J. (1995). *Lexical- Functional Grammar: A Formal System for Grammatical Representation. Formal Issues in Lexical-Functional Grammar*, ed. Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, Annie Zaenen, pp. 29-130. CSLI, Stanford.

Medero, J., Maeda, K., Strassel, S. and Walker, C. R. (2006). An Efficient Approach for Gold-Standard Annotation: Decision Points for Complex Tasks. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation.* 2450-3.

PARC. XLE. http://www2.parc.com/isl/groups/nltt/xle/