Whither Written Language Evaluation?

Ralph Grishman

Department of Computer Science
New York University
New York, NY 10003

Common evaluations have grown to be a major component of all the ARPA Human Language Technology programs. In the written language community, the largest evaluation program has been the series of Message Understanding Conferences, which began in 1987 [2,3]. These evaluations have focussed on the task of analyzing text and automatically filling templates describing certain classes of events. These conferences have certainly been a major impetus in the development of systems for performing such "information extraction" tasks, and thus in demonstrating the potential practical value of some of the written language processing technology.

There have been a number of concerns expressed, however, about the trend of these evaluations. First, these evaluations — and particularly the most recent, MUC-5 — have consumed large amounts of time, and in particular time spent learning and encoding detailed information about the domain, rather than learning about how to process language in general. Second, there has been a focus on technologies which are effective for this task but may not be effective for other, "language understanding" tasks.

In response to these concerns, a group of ARPA contractors and Government representatives met on December 2-4 in San Diego to plan for the next written language evaluation conference (MUC-6). This paper is a report on the conclusions of this meeting and some of the electronic interchanges which followed.

1. EVALUATION GOALS

Although the group met under the banner of MUC ("Message Understanding Conference"), it examined the issues of the evaluation of written language processing systems more generally, and did not limit itself to the types of evaluations conducted in past MUCs, which had been restricted to "information extraction" (template filling). The group began by considering the aims of such evaluations, which include

- assessing progress in written language understanding (and in particular, of ARPA's Tipster Phase 2 technology program)
- guiding research and pushing the technology (by identifying problems that need to be addressed)
- maintaining and increasing the interest and participation of potential users (by demonstrating systems which are "relevant" to practical applications)
- drawing more research groups into the evaluation process (and thus fostering the exchange of new ideas)
- lessening substantially the overhead associated with evaluations

To meet these various goals, the group proposed that MUC-6 consist of a menu of different evaluations. The evaluations would be run on a single test set, but there would be separate evaluation scores measuring different capabilities. Individual sites would be free to participate in any subset of the evaluations. (Of course, for sites which choose – or feel obligated – to participate to the maximum, the richness of the menu which was developed may work against the stated goal of reducing the evaluation overhead.)

The group decided that the corpus should consist of business-related articles from American newspapers and wire services. A large corpus of such texts, part of the corpora for the recent TREC (Text Retrieval Evaluation) Conferences, is available through the Linguistic Data Consortium. This includes articles from the Wall Street Journal, the San Jose Mercury News, and the AP newswire.

2. THE MENU

The menu of evaluations will include rather different types of tasks in order to meet the range of objectives cited above. On the one hand, we want to continue evaluation on tasks — such as "information extraction" — which can be seen as prototypes for real applications, and so will continue to draw interest from outside the natural language processing community. We would like to make these tasks as simple as possible, consistent with a semblance of reality, so that evaluation *per se* does not become a major time drain.

On the other hand, we are interested in exploring "glass box" evaluations — evaluations of the ability of systems to identify crucial linguistic relationships which we believe are relevant to a high level of performance on a wide variety of language understanding tasks. Of course, some people will believe that we have chosen the wrong relationships, or at least that natural language systems need not make these relationships explicit in the process of performing a natural language analysis task, and so will decline to participate in some or all of the glass box evaluations. We respect these disagreements, and have organized the menu of evaluations to take them into account. Any particular choice of internal evaluations necessarily represents some bet on the path of technical development. However, we believe that the relationships we have selected are sufficiently basic to understanding that the bet is worth taking, and that by encouraging work on these tasks we will push research on natural language understanding in ways which would not be possible with a limited application task such as information extraction.

The menu we came up with includes one task (named entity recognition) which is sufficiently basic to be characterized as both an internal and an application task; four internal evaluations; and two application-oriented evaluations:

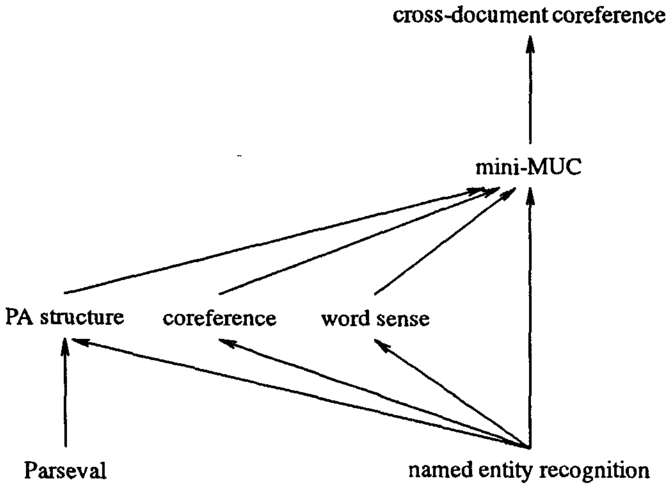

Figure 1: Potential interrelationships among the evaluations in MUC-6.

1. *Named Entity Recognition:* Identify company names, organization names, personal names, location names, product names, dates, times, and money.
2. *Parseval:* Bracket the syntactic constituents of the sentence.
3. *Predicate-Argument Structure:* Identify the relationship between lexical elements in terms of relations such as *logical-subject, logical-object*, etc.
4. *Word Sense Disambiguation:* Identify the word sense of each noun, verb, adjective, and adverb in the text, using the inventory of word senses from WordNet
5. *Coreference Resolution:* Identify identity of reference, superset, and subset relations among text elements, as well as situations where a text element is an *implicit* argument of another (e.g., a subject or object of a nominalization which appears elsewhere in the text).
6. *Mini-MUC:* Identify instances of a particular class of event in the text, and fill a template with the crucial information about each instance.
7. *Cross-Document Coreference:* Identify coreference relations between objects and events in different articles.

Evaluations 3, 4, and 5 are collectively known as *Semeval.* Each of the seven evaluations can be done independently, but there are potentials for using the results of the annotation for one task in performing another; these relationships are shown in Figure 1. Presumably, most participants will generate predicate-argument structure from parser output, so for them good Parseval performance would be a prerequisite for good performance on the predicate-argument metric. Recognition of named entities is essential for good performance on both Semeval and Mini-MUC. Some people will want to use the Semeval processing/output for the Mini-MUC, and some people won't; it is an interesting scientific question whether it helps. Cross-Document Coreference requires the output of Mini-MUC.

It will be possible for a site to investigate only one of these links, if they wished, rather than starting from the raw text input. This would allow people to build on others' work on named entity recognition, or to assess, assuming perfect or typical results on Semeval, how well one could do on Mini-MUC. Moreover, sites may be required to not only do a run using the (perfectly correct) key for the input to their component, but also using the (imperfect) actual results of some site participating in the full evaluation, which would be publicly available. (This might be arranged by staggering the evaluations, with the component evaluations scheduled before the mini-MUC evaluation.) These experiments would be analogous to the written-language-only part of the SLS evaluations.

3. THE EVALUATIONS

In this section we briefly describe each of the seven evaluation tasks. For each task we shall need to prepare a sample of text annotated with the information we wish the systems under evaluation to extract. To make the annotations more manageable and inspectable, we have combined the annotations for named entity recognition, coreference, and word sense identification. They are all encoded using an SGML tagging of the text, with separate attributes to record each type of information. Merging the annotations does not mean that the corresponding evaluations will be combined. We still expect that these three evaluations will be scored separately, and that text can be separately annotated for the three evaluations.[1]

To illustrate some of the annotations, members of the MUC-6 committee have annotated one of the "joint venture" news articles from the MUC-5 evaluation. The first two sentences of this article are:

> Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan. The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million New Taiwan Dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

The named-entity / coreference / word sense annotation is shown in Figures 2 and 3; the predicate-argument annotation is shown in Figure 4. All of these annotations are very preliminary; we expect they will be revised as annotation progresses.

3.1. Named Entity Recognition

The experience with MUC-5 indicated that recognition of company, organization, people, and location names is an essential ingredient in understanding business news articles, and is to a considerable degree separable from the other problems of language interpretation. In addition, such recognition can be of practical value by itself in tracking people and organizations in large volume of text. As a result, this evaluation may appeal to firms focussed on this limited task, who are not involved in more general language understanding.

In Figures 2 and 3, the named entity recognition is reflected in all the SGML elements besides *wd: entity* for companies and other organizations, *loc* and *complex-loc* for locations, *num* for numbers (including percentages), *date*, and *money*. Additional element types would be provided for other constructs involving specialized lexical patterns, such as times and people's names. For most of these elements, one of the attributes gives a normalized form: the decimal

[1] Although it will be simpler if at least the demarcation of named entities is performed first.

```
<s n=1>
<entity id=t1 type=company name='Bridgestone Sports CO' >                Bridgestone Sports Co. </entity>
<wd lemma=say sense=[verb.communication.0] >                             said </wd>
<date value='241189' >                                                   Friday </date>
<wd id=t2 identical=t1 >                                                 it </wd>
<wd >                                                                    has </wd>
<wd id=t3 sense=[verb.contact.0] >                                       set up </wd>
<wd >                                                                    a </wd>
<wd id=t4 sense=[noun.possession.0] >                                    joint venture </wd>
<wd >                                                                    in </wd>
<loc id=t5 name='Taiwan' type=country >                                  Taiwan </loc>
<wd >                                                                    with </wd>
<wd >                                                                    a </wd>
<wd id=t6 sense=[adj.all.0.territorial.0] args="[to t5]">                local </wd>
<wd id=t7 sense=[noun.group.0] >                                         concern </wd>
<wd >                                                                    and </wd>
<wd >                                                                    a </wd>
<wd sense=[adj.pert.0] >                                                 Japanese </wd>
<wd sense=[noun.relation.0] >                                            trading </wd>
<wd id=t8 sense=[noun.group.1] >                                         house </wd>
<wd >                                                                    to </wd>
<wd id=t9 sense=[verb.creation.0] args="[l-subj t4]" >                   produce </wd>
<wd id=t10 lemma=golf_club sense=[noun.artifact.0] >                     golf clubs </wd>
<wd >                                                                    to </wd>
<wd >                                                                    be </wd>
<wd lemma=ship sense=[verb.motion.0] >                                   shipped </wd>
<wd >                                                                    to </wd>
<loc id=t11 name='Japan' type=country >                                  Japan </loc>
<wd >                                                                    . </wd>
</s>
```

Figure 2: Named entity / word sense / coreference annotation of first sentence.

value of a number, a 6-digit number for dates, a standardized form for company names (following MUC-5 rules for company names).

3.2. Parseval

Parseval is a measure of the ability of a system to bracket the syntactic constituents in a sentence. This metric has now been in use for several years, and has been described elsewhere [1]. Parseval may eventually be supplanted in large part by the "deeper" and more detailed predicate-argument evaluation. However, for the present Parseval is being retained in order to accomodate participants focussed on surface grammar and participants reluctant to commit to predicate-argument evaluation until its design is stabilized and proven.

3.3. Predicate-argument structure

A very tentative predicate-argument structure for our two sentences is shown in Figure 4. As much as possible, we have tried to use the same structures which have been adopted by the Spoken Language Coordinating Committee for their predicate-argument evaluation. We summarize here, with some simplifications, only the most essential aspects of this representation.

For each event or state in the text, we introduce a Davidsonian event variable i, and treat the type and each argument of the event as a separate predication. So, for example, Fred fed Francis on Friday would be represented as[2]

(ev-type 1 eat)
(l-subj 1 Fred)
(l-obj 1 Francis)
(on 1 Friday)

Each elementary predication can be numbered by preceding it with a number and colon. Roughly speaking, a system would be scored on the number of such elementary predications it gets correct. Because this notation is none too readable, however, we also allow the abbreviated form

(eat [event 1] [l-subj Fred] [l-obj Francis] [on Friday])

where *[event 1]* could be omitted if there were no other references to the event variable. An entity, arising from a noun phrase with determiner *det* will be represented by

[2] Assuming that *on* is a primitive predicate, which is not expanded using an event variable. Otherwise we would have the predications (ev-type 2 on), (l-subj 2 1), and (l-obj 2 Friday).

```
<s n=2>
<wd >                                                                 The </wd>
<wd id=t21 sense=[noun.possession.0]identical=t4 >                    joint venture </wd>
<wd >                                                                 , </wd>
<entity id=t22 name='Bridgestone Sports Taiwan CO' type=company identical=t21 > Bridgestone Sports Taiwan Co. </entity>
<wd >                                                                 , </wd>
<wd id=t23 lemma=capitalize sense=[verb.cognition.1] >                capitalized </wd>
<wd >                                                                 at </wd>
<money id=t24 amount='20000000' unit='TWD' >                          20 million New Taiwan Dollars </money>
<wd >                                                                 , </wd>
<wd sense=[verb.stative.0] >                                          will </wd>
<wd sense=[verb.creation.1] >                                         start </wd>
<wd id=t25 sense=[noun.act.2] identical=t9 args"[l-subj t21] [l-obj t10]">  production </wd>
<wd >                                                                 in </wd>
<date id=t26 value='0190' >                                           January 1990 </date>
<wd >                                                                 with </wd>
<wd id=t27 sense=[noun.act.2] sub-of=t25 args="[l-subj t21]" >        production </wd>
<wd >                                                                 of </wd>
<num value='20,000' >                                                 20,000 </num>
<wd sense=[noun.artifact.1] >                                         iron </wd>
<wd >                                                                 and </wd>
<wd >                                                                 “ </wd>
<wd sense=[noun.artifact.0] >                                         metal wood </wd>
<wd >                                                                 ” </wd>
<wd id=t28 lemma=club sense=[noun.artifact.1] sub-of=t10>             clubs </wd>
<wd >                                                                 a </wd>
<wd sense=[noun.time.0] >                                             month </wd>
<wd >                                                                 . </wd>
</s>
```

Figure 3: Named entity / word sense / coreference annotation of second sentence.

e: (det <restr1 restr2 ...>)

Each $restr_i$ is a constraint on the entity, stated as a predication on index e. Thus "the brown cow which licked Fred" would be represented by

1: (the <(brown [l-subj 1]) (cow [l-subj 1])
(lick [l-subj 1] [l-obj Fred])>)

The notation "$?i$" means that i is optional; the notation i / j means that either i or j is allowed.

The written language group, however, is not taking the same approach to the selection of predicates and role-names as the spoken language group. The spoken language group aspires to a truly semantic representation, independent of the particular syntactic form in which it was expressed. This seems feasible in the highly circumscribed domain of air traffic information. It does not seem a feasible near-term goal for all of language, or even for all of "business news", which is a very broad domain. Instead we will be initially using a form of grammatical functional structure, with lexical items as heads (predicate types), and role names such as *logical subject* and *logical object*. The representation will be normalized with respect to only a limited number of syntactic alternations, such as passive, dative with "for", and dative with "to". I expect that the representation will gradually evolve to normalize a larger number of paraphrastic alternations.

3.4. Coreference

Coreference can be annotated either at the level of the word sequence or at the level of predicate-argument structure. By recording coreference at the word level, we lose some distinctions that can be captured at predicate-argument level. On the other hand, annotating at the word level allows for evaluation of coreference without generating predicate-argument structure. So — in order to keep the menu items as independent as possible — our current plan is to annotate coreference at the word level, with the head word of the anaphor pointing to the head word of the antecedent.

Coreference is recorded through attributes in the SGML annotation (Figures 2 and 3). For purposes of reference, elements are annotated with an *ident* attribute. Identity of reference is indicated by an attribute *identical* pointing to the antecedent. A superset/subset relation is indicated by a *sub-of* attribute. Finally, if a predication has implicit arguments which are coreferential with prior text elements, they are annotated as *args = "[role antecedent]"*.

```
(DECL <(say [event 1]
         [l-subj 2:Bridgestone-Sports-Co.]
         [l-obj  <(set-up [event 3]
                    [l-subj 4|2]
                    [l-obj 5:(a <(joint-venture [l-subj 5])
                                 (in [l-subj 5] [l-obj  Taiwan])
                                 6:(with
                                     [l-subj 5|3]
                                     [l-obj  7:(ANDNP <8:(a <(concern [l-subj 8])
                                                             (local [l-subj 8])>)
                                                       9:(a <(trading-house [l-subj 9])
                                                             (Japanese [l-subj 9])>)>)])
                                  10:(PURPOSE
                                       [l-subj 5|3]
                                       [l-obj <(produce
                                                 [l-subj ?5]
                                                 [l-obj 11:(NO-DET <(golf-club [l-subj 11])
                                                                   (PLURAL [l-subj 11])
                                                                   (PURPOSE
                                                                     [l-subj 11]
                                                                     [l-obj <(ship [event 12]
                                                                                [l-subj ?5]
                                                                                [l-obj 11])
                                                                              (to [l-subj 12]
                                                                                [l-obj Japan])
                                                                                >])>)])>])>)])
                    ?6 ?10
                    (PERFECT-TENSE [l-subj 3])>])
       (PAST-TENSE [l-subj 1])
       (AT-TIME [l-subj 1]
                [l-obj  13:(DATE <(DAY-OF-WEEK [l-subj 13] [l-obj Friday])>)])>)


(DECL <(start [event 1].
              [l-subj 2:(the <(joint-venture [l-subj 2])
                              (IDENTICAL [l-subj 2]
                                         [l-obj Bridgestone-Sports-Taiwan-Co.])
                              (capitalize [event 3] [l-obj 2])
                              (at [l-subj 3]
                                  [l-obj 4:(NO-DET <(New-Taiwan-Dollar [l-subj 4])
                                                    (PLURAL [l-subj 4])
                                                    (CARDINALITY [l-subj 4]
                                                                 [l-obj  20000000])>)])>)]
              [l-obj 5:(NO-DET <(produce [event 5] [l-subj 2])>)])
         (FUTURE-TENSE [l-subj 1])
         (in [l-subj 1]
             [l-obj 6:(DATE <(MONTH [l-subj 6] [l-obj January])
                             (YEAR  [l-subj 6] [l-obj 1990])>)])
         (with [l-subj 1]
               [l-obj 7:(NO-DET <(produce [event 7]
                                   [l-subj 2])
                                   [l-obj  8:(NO-DET <(club [l-subj 8])
                                                      (PLURAL [l-subj 8])
                                                      (and <(iron [l-subj 8])
                                                            (metal-wood [l-subj 8])>)
                                                      (PER [l-subj (CARDINALITY [l-subj 8]
                                                                                [l-obj 20000])]
                                                           [l-obj month])>)])>)])>)
```

Figure 4: Predicate-argument structure.

3.5. Word sense identification

The third element of the Semeval triad is sense identification. As a sense inventory, we haved used WordNet, which is widely and freely available and is broad in coverage [4]. The notation used to refer to particular WordNet sense was described in [5].

3.6. Mini-MUC

This component is the direct descendant of the information extraction tasks in the previous MUCs [2,3]. In response to criticism that the evaluation task had gotten too complex, we have endeavored to make the new information extraction as simple as possible. The template will have a hierarchical structure, as in MUC-5, but probably with only two levels of "objects". The objects at the lower level will represent common business news entities such as people and companies. A small inventory of such objects will be defined in advance. The upper level object will then be a simple structure with perhaps four or five slots, to capture the information about a particular type of event.

The following were suggested as typical of such templates:

1. Location of use of pollution control products.
Product:
Purchaser:
Act:
Act-Location:

2. Org. ordering or cancelling order for aircraft.
Manufacturer:
Model:
Buyer:
Order status:

3. Companies quote prices on products.
Company:
Products:
Prices:
Date:

4. Stock market min/max during interval.
Market:
Index:
Extreme: H/L
Epoch:

Because of the simplicity of these templates, a month was felt to be sufficient time for developing a particular extraction system. In fact, because of concern that a single template introduced too much risk due to possible faulty template/problem design, it was suggested that working on three closely related topics within a single month might be desirable.

3.7. Cross-document coreference

One way in which prior MUC tasks were unrealistic is that they did not attempt to link events across documents, even though the corpus frequently included multiple documents about the same event. To remedy this shortcoming, it was suggested that the task of making such event coreference links across documents be included as an additional item on the evaluation menu.

4. PLANS

The menu of evaluations which has been developed for MUC-6 is certainly ambitious; perhaps it is too ambitious and will need to be scaled back. While the cost of participating in a single one of these evaluations should be much less than the effort required for MUC-5, the effort to prepare all these evaluations will be considerable. Detailed specifications will need to be developed for each of the evaluations, and substantial annotated corpora will have to be developed, both as the "case law" for subsequent evaluations and as a training corpus for trainable analyzers. If this is all successful, however, it holds the promise for fostering advances in several aspects of natural language understanding.

A description of the menu of evaluations was disseminated electronically at the end of December 1993. Further details, including a sample annotated message, were distributed at the end of February 1994. After a period of public electronic comment, we shall be recruiting volunteer sites to begin annotating texts, slowly over the course of the spring, as the specifications are ironed out, more rapidly over the summer, once specifications are more stable.

A dry run evaluation, possibly including only a subset of the menu items, will be conducted in late fall of 1994; MUC-6 is tentatively scheduled for May of 1995.

5. Acknowledgements

I wish to thank Jerry Hobbs for sharing with me his write-up of the MUC-6 meeting, and to thank the people who prepared the individual annotations: Jerry Hobbs for predicate-argument and coreference; George Miller and his colleagues at Princeton for word senses; and Beth Sundheim for named entities. Finally, I wish to thank all the participants in the MUC-6 meeting — Jim Cowie, George Doddington, Donna Harman, Jerry Hobbs, Paul Jacobs, Boyan Onyshkevych, John Prange, Len Schubert, Bill Schultheis, Beth Sundheim, Carl Weir, and Ralph Weischedel — for their contributions.

The author was supported in the preparation of this report by the Advanced Research Projects Agency under Grant N00014-90-J-1851 from the Office of Naval Research.

References

1. Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., and Strzalkowski, T. A procedure for quantitatively comparing the syntactic coverage of English grammars. *Proc. Fourth DARPA Speech and Natural Language Workshop*, Feb. 1991, Pacific Grove, CA, Morgan Kaufmann.
2. *Proceedings of the Third Message Understanding Conference (MUC-3)*. Morgan Kaufmann, May 1991.
3. *Proceedings of the Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufmann, June 1992.
4. Miller, G. A. (ed.), WordNet: An on-line lexical database. *International Journal of Lexicography* (special issue), 3(4):235-312, 1990.
5. Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T., "A semantic concordance", Proc. Human Language Technology Workshop, 303-308, Plainsboro, NJ, March, 1993, Morgan Kaufmann.