Summary of Workshop on Lexicons for Text Extraction

James Pustejovsky
Computer Science Department, Brandeis University
Email: jamesp@cs.brandeis.edu

This workshop discussed the problems with lexicon development in the context of MUC-style application programs. The topics ranged from general issues in lexicon portability (Cahill), to Japanese lexicons (Mauldin) and problems encountered with MRDs in sublanguage domains (Pustejovsky).

The POETIC Lexicon Lynne Cahill, of the University of Sussex, England, presented the design and specifications for the lexicon used in their Traffic Information Collator system, and what problems they encountered in porting it to the MUC task. This is an information extraction system used by local police personnel for traffic reports. The domain is characterized by a fairly tight and limited vocabulary, as well as a telegraphic style of syntax. Cahill discussed the relative ease with which the lexicon was adapted to new police force domains.

The general issues raised in Cahill's presentation were the problem of going from a sublanguage lexicon to a broader lexicon, as required for the MUC-5 English Joint Venture domain. The porting effort took 12 person months in 6 months time.

The MUC-5 lexicon design consists of a domain specific lexicon and a phrasal lexicon. These were used in conjunction with the Alvey Natural Language Toolkit for parse recovery. Rich lexical information was added only to words which were significant in the domain as triggers for the template fills. Furthermore, the recognition of company and personal names was accomplished by standard pattern matching techniques.

Cahill discussed the different nature of the lexical entries in the two domains. Porting to MUC-5 required a new semantics and much more syntax. The result was that the incomplete lexicon gave rise to undergeneration of appropriate template objects, while fragmentary parsing resulted in template overgeneration, because of the liberal acceptance of too many patterns. There was, furthermore, no contextual feedback into the parser, as well as no way of selecting the most likely analysis of a given pattern, if several fired. Cahill pointed out that these problems were largely due to time constraints in the development cycle, rather than the nature of the lexical design.

Lexical Information in SHOGUN Michael Mauldin, of CMU, presented information about MAJESTY, the Japanese lexicon in SHOGUN. This lexicon contains over 17,000 entries, including open class words, proper names, locations, and numeric entries.

Mauldin's talk focussed on the parts of speech and Japanese segmentation using both the MAJESTY and JUMAN programs, and the use of the JUMAN to MAJESTY converter. The size of the Japanese lexicon is: 13,892 completed word entries, 17,943 word senses.

Mauldin then compared the JUMAN and MAJESTY parts of speech, and the segmentation agreement between MAJESTY and JUMAN using the converter. He found that segmentation agreement between the two was 83.2%, while segmentation and POS agreement was 76.9%.

CMU has made this lexicon a shareable resource, and it is available by anonymous FTP from CMU at the following location:

```
Host: nl.cs.cmu.edu
Dir: /usr/mlm/ftp/tipster
```

The files available are:

```
jlex.tar.Z          (Japanese lexicon)
jjv-seg.tar.Z       (Segmented JV corpus (by MAJESTY))
jap-industry.rules  (Rules for Japanese SIC codes)
j2m.tar.Z           (JUMAN -> MAJESTY converter)
name-acq.tar.Z      (Japanese Name Acquisition s/w)
```

Limitations of MRDs and Sublanguage Lexicons In the last presentation, James Pustejovsky of Brandeis University discussed the inherent limitations of extracting information from machine-readable dictionaries, and made the observation that there is an inverse correlation between the utility of *direct* MRD-derived lexical items with the tightness of a sublanguage.

From Pustejovsky's experience in the NMSU/Brandeis Tipster effort, where domain lexicons were semi-automatically seeded from an MRD-derived core lexicon (LDOCE), problems arose with the usefulness of some MRD data. Because some sublanguage senses for key (trigger) lexical items are so removed from dictionary senses, sense determination and acquisition must come from domain-specific corpora.

Pustejovsky presented the dimensions along with to analyze the usefulness of MRD fields:

- Categorization of the word for tagging;
- Subcategorization variants; transitive or intransitive;
- Semantic category of the word (sense identification), and semantic type of subcategorized elements.

The problems with direct MRD-lexicons for sublanguages can be summarized as follows:

- Category distribution specified in MRD may not reflect the actual use of the word in the corpus or domain;
- Subcategorization variants are weak indicators of the actual syntactic variation in the corpus; i.e. forms that are not encoded in the MRD;
- Meaning shifts occur in sublanguages that are not encoded in the MRD.

Pustejovsky then turned to evaluation issues and how they relate to lexicon development. If we were to be interested in general word sense identification and predicate-argument structure in the text, and not just differentiating trigger terms from free text, the style of lexicon development would be very different. Some sort of core lexicon would be very useful as a common resource from which to tune to specific domains and tasks, through statistical corpus techniques. In fact, even the sublanguage use of general vocabulary items is predictable only from examination of the corpus. Thus, corpus acquisition and tuning becomes an integral part of any lexical system development.