AN INTERPRETATIVE DATA ANALYSIS OF CHINESE NAMED ENTITY SUBTYPES

Thomas A. Keenan
Department of Defense,
9800 Savage Road,
Fort Meade, Md. 20755
tomkeena@romulus.ncsc.mil

1. MOTIVATIONS FOR AN INTERPRETATIVE DATA ANALYSIS

"In assessing the performance of information extraction systems, we are interested in knowing the classes of errors made and the circumstances in which they are made."[1] However, to date the Tipster scoring categories (*correct, partial, incorrect, spurious, missing*, and *noncommittal*) have not been applied to classes of data based on structural distinctions in the language, or on semantic subclasses more finely differentiated than the NE types (*person, location, organization, time, date, money,* and *percent*). For example, there has been no attempt to score the extraction of transliterated foreign person names, or of short-form aliases of corporation names, or of Julian dates as opposed to Gregorian dates as opposed to dates of the Chinese lunar calendar.

There are obvious practical reasons for this. The scoring criteria are limited to those that can be measured without access to anything more than the annotations the systems generate [2], and those applied by human taggers to the answer keys. Moreover, any new annotations that might become available represent a limited subset of the infinite number of ways that NE data might be subcategorized, in accordance with particular interests, applications, and capabilities.

Yet, from among these innumerable possible subcategories of entity names, a few would seem likely to emerge as more well-motivated than the rest. Note that an appendix on "VIP Names" or "Country and Capital City Names" is more likely to appear in a desk-top dictionary than a list called "Ethnic Surnames in their Native Scripts and Common Anglicized English Renderings." One would expect observant end-users of information extraction systems to notice rather quickly that certain high frequency, hard-to-get, or thematically significant categories of names are missing or incorrect in the output. And one might desire, at some point in the system development loop, to capture these observations systematically, so as to direct efforts at system improvement. This would be especially desirable if system development includes relatively labor-intensive linguistic analysis.

Doing this "systematically" is not the same as measuring errors scientifically. To count the number of tagged VIP person names, for example, presupposes somebody's interpretation of whether "VIP" includes only chiefs-of-state, or chiefs-of-state and cabinet ministers, or these plus nobel prize winning scientists, novelists, peace activists, etc. So, the following observations are at best an *interpretative* error analysis, informed by knowledge of the language and of likely user expectations. However, we try to define this as a series of steps that reasonably approximates a scientific discovery procedure.

2. A PROCEDURE FOR INTERPRETATIVE DATA ANALYSIS

The following steps were taken to analyze the MET Chinese test data.

- Step 1: Scan the Input Data for Salient Subtypes

The MET Chinese named-entity data tagged in the test keys was scanned for sub-classes of names appearing to meet one or more of the following criteria:

- they are *frequently-occurring* semantic subtypes (for example it was apparent that country names comprised a large subclass within the LOCATION-tagged data);

- they are *readily listable, high-interest* subtypes (for example, "chiefs-of-state" comprise a class small enough to be readily listable by a human analyst; in addition, we might expect their activities to be more newsworthy than those of the newspaper reporters or official government spokespersons whose names also appear frequently in the data).

- they are *readily patternable subtypes* (for example, many taggable organization names begin with a location name and end with a unit designator, as in the name "*Minnesota* Mining and Manufacturing *Corporation*". Other organization names, such as "Hamas," had no obviously specifiable morphological features in common with large numbers of other names.

This scanning process identified a small number of data subtypes, which were individually describable in terms of the meaning, forms or distributions of names, and which collectively seemed to comprise a large percentage of all names extracted [3]. The resulting inventory of subtypes can be thought of as an ***hypothesis*** that the NE data is describable in a certain interpretative yet systematic way.

- Step 2: Count NE Occurrences by Subtype

Tagged names were searched by NE "type" (person / location / organization) using a concordance tool (NMSU's "XConcord"), then copied to files representing each of the posited subtype classes, or to a catch-all "residual" class. The number of names in each file was then counted to arrive at an overall profile of the data distribution. This step can be thought of as a ***test*** of the data distribution hypothesis.

- Step 3: Chart the Distribution of NE Data

Table 1 (following the text) provides a ***summary of the "test" results.***

- Step 4: Check for Inconsistencies in the Data Distribution

The numbers in the boxes of Table 1 were tallied and analyzed for internal consistency and non-conformity to our original expectations, that is, to show that the "hypothesis" was not invalidated. If no inconsistencies were found and an acceptably high percentage of the data had been accounted for, then the descriptive category set might have appeared adequate. Note, however, that the ratio of "residual" person names, 40%, is considerably higher than the ratios of residual location and organization names. This suggested that the initial description was leaving a significant portion of the data unaccounted for.

- Step 5: Loop Back to Step 1 (or stop when an acceptably high percentage of the data is accounted for, and inconsistencies are resolved)

Re-examination of the data revealed that, among the 2-3 syllable, non-VIP, "residual" person names, 40% [4] are directly preceded on the left by a "title" (e.g. "Representative so-and-so"). This still leaves 24% [5] of person names unaccounted for. Some high percentage of this 24% presumably could be accounted for by an adequate structural description of Chinese "surname plus given name" patterns. This description, and measurement of the data it would cover, was not attempted, due to time constraints and the complexity of the problem.

3. APPLICATIONS OF THE INTERPRETATIVE DATA ANALYSIS

As suggested above, variations of the above procedure can be used to generate profiles of the data in order to direct efforts at system improvement. This may or may not be worth the cost of analysis if system improvement is driven solely by piping more and more massive amounts of development data into a statistical learning engine. If sufficiently massive and varied development data is available, presumably the system eventually will train upon something approaching all of the relevant data subtypes, without any need to know and describe what those subtypes are. However, when the approach involves labor-intensive pattern development based on linguistic structures, future language-analytic development could be focused by applying in advance something like the foregoing procedure, supported by tailored versions of concordance tools and other on-line analytic aids.

4. FOR FURTHER ANALYSIS

Under what circumstances would this approach be useful? Above it was observed, "we are interested in knowing the classes of errors made and the circumstances in which they are made." We might also be interested in knowing the *significance* of different classes of errors, or of data correctly handled compared to data incorrectly handled. Apparently only a small percentage of taggable NE's fall into the "residual" subclasses (see Table 1), and these tend to correspond to data that is "hard to get" for the MET participants [6]. Although the percentage may seem small, the signifi-

cance of a 10% or even 5% error rate ultimately depends on the usefulness of the unexploited data to the end-user. If "hard-to-get," "residual," low-frequency data tends to correspond to certain semantic subclasses and not others, then the possibility arises that these classes of NE's assume some non-random thematic significance in the texts. That is, obscure, low-frequency proper names (or names not conforming to prevailing patterns and types) might be expected to crop up in newspaper texts only when they become highly newsworthy, and yet be precisely the kinds of names that some systems tend to miss. It remains for further analysis to test and measure for these possibilities by observing system performances on specific subclasses of data selected for textual significance.

5. REFERENCES

[1] Sundheim, Beth M. "Tipster/MUC-5 Information Extraction System Evaluation" in Proceedings of the Tipster Text Program (Phase I), p. 149. September 1993. Morgan Kaufmann Publishers, Inc. San Francisco.

[2] ibid.

[3] *Zipf's Law* predicts that a relatively small number of high-frequency words will account for a disproportionately large number of token occurrences in a text corpus. A similar generalization appears to govern language patterns as well as words, that is, a small number of high-frequency patterns apparently can account for a disproportionately large share of types such as organization names, with the specification of new patterns eventually reaching a point of ever-diminishing returns.

[4] i.e. 40% X 40% = 16% of all person names

[5] 40% - 16%

[6] This finding is based on the author's preliminary, unpublished analysis. In general, MET participant programs appear to be most successful in tagging subclasses of data covered under the "listable" column of the following chart, and least successful against the "residual" subclasses.

All NE's: 3100	"LISTABLE"s	"PATTERNABLE"s	"RESIDUAL"
All LOC's: **1700** (50%)	**unabbreviated Country/ Major City Names:** 80% of LOCs (45% of NEs)	**abbreviated Country/ Major City names:** 15% of LOCs (10% of NEs)	**"obscure" LOCs (not country/major city names):** 5% of LOCs (~0% of NEs)
All ORG's: **700** (25%)	**Major International & Chinese ORGs:** 35% of ORGs (10% of NEs)	**LOC... "UNIT" pattern:** 50% of ORGs[a] (10% of NEs)	**other not readily patternable ORGs:** 5% of ORGs (~0% of NEs)
All PER's: **700** (25%)	**VIPs** (Chiefs of State, etc.): 15% of all PERs (5% of NEs)	**3+ syllable non-VIP foreign PERs:** 40% of PERs (10% of NEs)	**2-3 syl., non-VIP, "unbound" PERs:** <40% of PERs (10% of NEs)*
		2-3 syl., non-VIPs, bound by punctuation & "said": 5% of PERs (~0% of NEs)	* *about 40% of the above category are bound on the left by a "title" (e.g. "Representative so-*and-so")
TOTAL	60% "listable"	30% "patternable"	10% "residual"

Table 1: Chinese NE Data Distribution (figures rounded to nearest 5%)

a. Approximately 10% of all ORGs appear to fall into small listable/patternable classes not indicated