Using Statistics Gained From Corpora in a Knowledge-Based NLP System

Doug McKee and **John Maloney**
SRA
2000 15th Street North
Arlington, VA 22201
mckeed@sra.com, maloneyj@sra.com

Abstract

Knowledge acquistion is a bottleneck for broad-coverage knowledge-based natural language processing systems. Statistical techniques can help by providing automatic knowledge acquisition as well as helping to focus manual acquisition effort. In the following paper we discuss how we automatically acquire syntactic and semantic knowledge from a corpus for use in the SOLOMON natural language understanding system.

Introduction

For the past several years, SRA has been developing a knowledge-based NLP system, SOLOMON, which has been used within several large-scale data extraction applications. Domains have included financial and medical texts, for which SOLOMON was successfully ported, and the system is also being used in Spanish and Japanese data extraction. [Kehler *et al.*, 1990; Shugar *et al.*, 1991].

Along with many in the field, we have found that the knowledge acquisition needed to extend a knowledge-based system is quite onerous. This motivated us to investigate the use of statistical techniques to reduce the level of effort previously necessary to accomplish this. We have found a variety of areas where such techniques provide valuable assistance to a knowledge-based system. These include: (1) establishing the special features of new domain language that have to be dealt with; (2) generation of lexical, syntactic, and semantic information from corpora using statistical techniques; and (3) using statistical knowledge to aid processing directly when hand-coded information is sparse.

So far, our methods have required very few changes to the SOLOMON processing modules. We have investigated the application of statistical techniques in several of the modules (see next section for a system description), but have not seen any need to modify SOLOMON's basic knowledge-based approach. We instead use statistical techniques where they help to produce knowledge for use by SOLOMON, and where they can increase the efficiency of processing.

We will first present a description of SOLOMON as a basis for what follows. Then we discuss the manner in which we use statistical techniques to specify the characteristics of new domain language. Last, we cover our work in statistically based syntactic and semantic knowledge acquisition, as well as the role of that knowledge in actual processing.

SOLOMON

There are four processing modules of the SOLOMON system: Preprocessing, Syntactic Analysis, Semantic Interpretation, and Discourse Analysis.

The Preprocessing module takes as input a list of words and performs word lookup, morphological analysis, and preparsing. The latter includes looking for domain multi-words and extremely common, but hard to parse, phrases such as dates or proper names. SOLOMON uses a range of lexicons, including a general vocabulary and specialized domain lexicons. Included in each lexicon entry is a word's linguistic meaning and a pointer to a corresponding concept in a knowledge base. While the lexicons contain language-dependent linguistic information about words, the knowledge bases hold language-independent information about the world. In fact, SOLOMON uses the same knowledge bases to extract data from multilingual texts.

The Syntactic Analysis module takes the output of Preprocessing and puts it through general parsing, and if necessary, applies a more robust "debris parsing" mechanism. General parsing uses an implementation of Tomita's algorithm [Tomita, 1986], modified to use augmentations on grammar rules. The grammar is a set of context-free phrase structure rules, augmented with context-sensitive constraints as well as routines to build a structure akin to a Lexical-Functional Grammar (LFG) f-structure. In addition to performing syntactic checks, the rule constraints also do coarse-grained semantic checks. Constituents are "weighted" as they are built based on the weights of their sub-constituents and the particular phrase structure rule. The augmentations for less commonly fired grammar rules assign worse weights to the constituents they

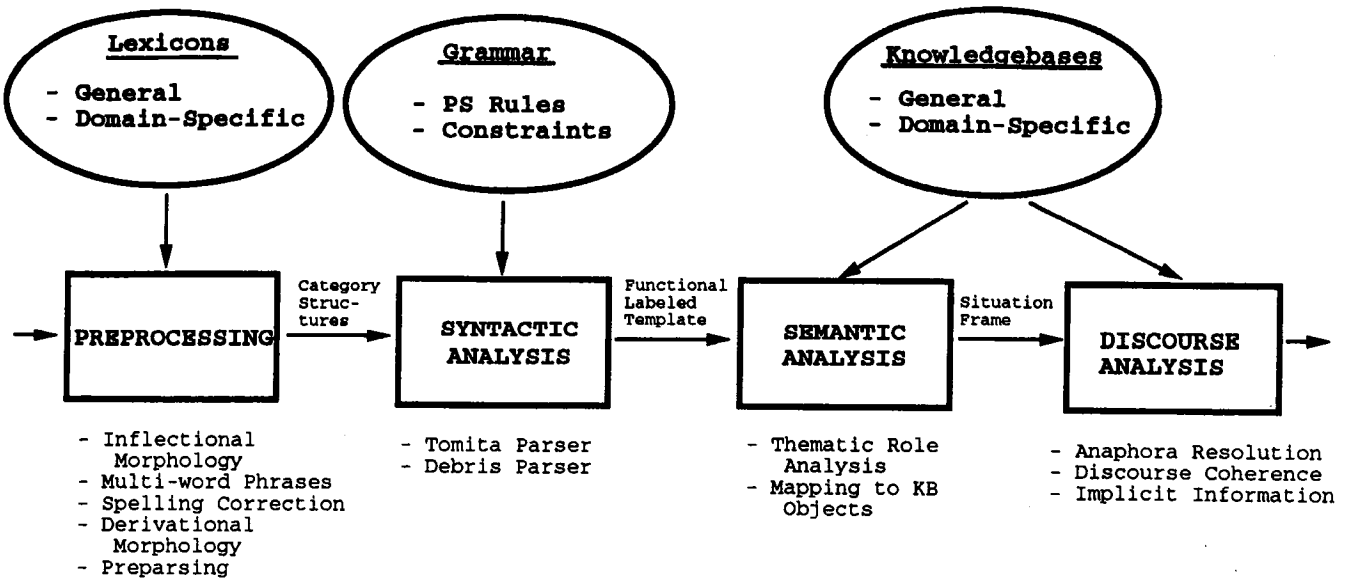

Figure 1: The SOLOMON System Architecture

build. Leaf constituents (i.e., the lexical items) contain weights in the lexicon.

When a constituent passes below a certain weight threshold, it is judged too unlikely and that grammar rule is not applied. After parsing, we sort all the possible resulting constituent structures by weight and send the best one to Semantic Interpretation.

The Semantic Interpretation module is compositional off of the syntactic structure and uses a thematic role analysis in building a "Semantically-Labelled Template." Predicate-argument relations and selectional restrictions for verbs are encoded in predicate classes in the knowledge base, with only "idiosyncratic" information stored in the lexicon. In addition, this information is accessed during parsing to make disambiguation decisions and to prune anomalous parses.

Discourse Analysis performs focus-based reference resolution and fills in implicit information where necessary. The resulting interpretation is then mapped to a final representation in a knowledge representation language.

Bounding the Domain

When porting SOLOMON to a new domain, the first step is to specify the domain's unique linguistic features and the special world knowledge required. This includes domain vocabulary, phrases, syntactic constructions, as well as typical domain events and scenarios ("scripts"). In the past, we have found that this process is laborious and too subject to bias when done entirely by hand. Statistical techniques offer a viable alternative which can capture the full range of linguistic variation and novelties found in a new set of domain texts.

The first step is to identify those words in a new domain corpus which are most likely to have domain meanings. To do this, we compare the frequencies of words in domain texts with frequencies of the same words in other texts. In our current work for the Message Understanding Conference (MUC) 4, we began by comparing the MUC corpus with a 2.7 million word corpus of Dow Jones news articles from the Penn Treebank. Our MUC corpus consisted of the 1300 development texts from the Central and South American terrorism domain. The Dow Jones corpus consists of articles slanted towards business and finance. We have found that comparing the differences in frequencies is superior to just taking the most frequent words in a corpus in determining the set of domain words.

Figure 2 shows the 30 open-class words in each corpus with the greatest frequency differences. This number is calculated by subtracting the frequency of the word in the Dow Jones corpus (i.e., number of occurrences of the word divided by the total numbers of words in the corpus) from that same word's frequency in the MUC corpus. Figure 2 contains intuitively good results for both corpora.

Another important step is to set the lexical weights on word senses in the lexicon, which is of crucial importance when new domain items are introduced that might be homonymous with general-lexicon words. These weights are used by SOLOMON during pars-

MUC word	Frequency Difference	Financial word	Frequency Difference
SALVADOR	0.0035	SAID	-0.003
FMLN	0.0031	COMPANY	-0.0028
GOVERNMENT	0.003	NEW	-0.0023
SAN	0.0029	YEAR	-0.0022
SALVADORAN	0.0029	SAYS	-0.0021
FORCES	0.0026	MARKET	-0.002
PEOPLE	0.0025	STOCK	-0.0017
ARMED	0.0025	SHARE	-0.0013
NATIONAL	0.0019	SHARES	-0.0013
EL	0.0018	TRADING	-0.0011
COUNTRY	0.0017	SALES	-0.0011
POLICE	0.0017	MR.	-0.001
ARMY	0.0017	CORP.	-0.001
TODAY	0.0014	BUSINESS	-9.0E-4
ATTACK	0.0013	COMPANIES	-9.0E-4
PEACE	0.0012	INC.	-9.0E-4
PRESIDENT	0.0012	YORK	-9.0E-4
GUERRILLAS	0.0012	PRICE	-8.0E-4
POLITICAL	0.0012	FEDERAL	-8.0E-4
FRONT	0.0012	BANK	-8.0E-4
CRISTIANI	0.0011	PRICES	-8.0E-4
KILLED	0.0011	QUARTER	-7.0E-4
DRUG	0.0011	CO.	-7.0E-4
RADIO	0.0011	SECURITIES	-7.0E-4
LIBERATION	0.0011	EXCHANGE	-7.0E-4
BOGOTA	0.0011	INVESTORS	-7.0E-4
TERRORIST	0.001	CENTS	-7.0E-4
COLOMBIAN	0.001	CHAIRMAN	-6.0E-4
STATES	9.0E-4	TAX	-6.0E-4
REPORTED	9.0E-4	RATE	-6.0E-4

Figure 2: The 30 most domain-specific open-class words based on frequency comparisons for two domains

ing to choose among homonyms based on likelihood of occurrence. For example, we used the Dow Jones corpus to weight words with cross-categorial ambiguity when building a system to process text from the financial domain.

The third step in bounding the domain is to automatically identify very common domain phrases and "multi-words." The latter are phrases that act as one unit and are generally noncompositional, e.g., United States of America.

Multi-words, if sent through unanalyzed, cause unnecessary ambiguity in parsing. Our statistical method to identify the sequences of words that constitute multi-words is an algorithm based on Fano's mutual information measure [Fano, 1961], later used by Church in his work [Church and Hanks, 1990] as a Mutual Association Ratio for words in text. Our algorithm determines those sets of words which are directly adjacent with a higher than chance probability. Figure 3 shows the 10 most probable multi-words along with their components' Mutual Association Ratio occurring in a set of medical articles taken from the set of information retrieval test texts made available by Ed Fox via anonymous ftp.

Once identified, these multi-words are automatically transformed into suitable lexicon entries. Since the majority of these items behave as nouns, they are classified that way in the lexicon. We are currently extending our algorithms to recognize longer multi-word phrases using techniques similar to those of Smadja and McKeown [Smadja and McKeown, 1990].

Automatically Acquiring Syntactic Knowledge

We have used statistical techniques in various ways to assist SOLOMON in arriving at the best syntactic analysis of sentences. To derive full accurate parses, a knowledge-based system needs to have information about the syntactic properties and preferences of verbs. We have therefore concentrated our efforts on determining what kinds of constituents frequently occur in the vicinity of verbs. Using this data, we can establish the verb's attachment preferences. We have automatically derived verb subcategorization frames from corpora, as well as information on verb transitivity. Such statistically acquired knowledge has increased the accuracy of SOLOMON's parsing.

We have found that domain-specific syntactic knowledge can be acquired by running our algorithms over often small domain corpora. An interesting comparison is with Brent's work [Brent, 1991]. He uses untagged text and extracts syntactic information from it by looking at unambiguous cases. While such an approach yields very accurate results, it requires a very large corpus, which is frequently not available for specific domains. We have used similar algorithms on text tagged for part of speech that work well on much smaller corpora, because the tagging allows identification of many more unambiguous cases than Brent's work.

Prepositional preferences: We can often detect subcategorized prepositions statistically because they appear much more frequently than expected after verbs, as Hindle and Rooth [Hindle and Rooth, 1991] have shown most recently. Using a simpler algorithm, we get very similar results. In our algorithm, a verb subcategorizes for a given preposition if the Mutual Association Ratio between the verb and the preposition exceeds 2.0, where the permissible "window" is two (i.e., a maximum of one word occurring between the verb and preposition). We have applied these same techniques to domain texts and discovered that subcategorization information can vary depending on domain. Figure 4 shows several verbs and the prepositions they subcategorize for based on our processing of the MUC and Dow Jones corpora. SOLOMON uses this information during parsing, strongly preferring attachments that allow a verb to subcategorize.

Clausal attachment By examining unambiguous cases of clause attachment in free text, we can learn what verbs take what types of clauses as arguments. We distinguish between "that" complements (THATCOMPS), infinitive complements (TOCOMPS), and gerund complements (INGCOMPS). [1] Figure 5 shows a sample of verbs and what types of clauses our algorithms determined will attach to them. Note that "eat" takes no clausal arguments while "want" takes a TOCOMP.

Knowing the clausal subcategorization of verbs reduces parsing ambiguity. It is desirable from the standpoint of parsing efficiency to prefer the known subcategorizations in making attachment decisions.

Transitivity rating: We use coocurrence information between verbs and directly following noun phrases to determine the transitivity of verbs. For example, in "Mary [verb] John ..." it is likely that [verb] is transitive. The transitivity rating of a verb is defined as the number of transitive cases (i.e., where there is a directly adjacent surface object or the verb occurs in a passive construction) divided by the total number of

[1] We define "unambiguous cases of clause attachment" in the following way: *that* tagged as a subordinating conjunction directly following a verb; *to* and the base form of a verb directly following a verb; a verb tagged as a gerund directly following a verb. Identifying the differences between "I live to eat" (a reduced "in order to" clause) and "I want to eat" (a real infinitive clause) turns out not to be necessary. Even though both look syntactically similar, verbs that take real infinitive complements are much more often found in that configuration; "in order to" clauses are not statistically frequent.

Domain phrase	Mutual Association Ratio
CARBONIC ANHYDRASE	15.7
LIVEDO RETICULARIS	15.7
HYDROGEN PEROXIDE	15.7
VENA CAVA	15.7
OBTUSA MCCOY	15.7
UNITED KINGDOM	15.2
EMISSION SPECTROGRAPHY	15.2
SUICIDAL TENDENCIES	15.1
THERMAL DENATURATION	15.1
SPINA BIFIDA	14.7

Figure 3: Ten of the most likely two-word phrases from a set of 1033 medical articles

word	MUC prepositional preferences	Dow Jones prepositional preferences
negotiate	with	with
accept	as	–
confront	with	with
decide	on	–
talk	over, with	with
stick	–	out, with
prevent	–	from

Figure 4: Seven verbs and the prepositions they probably subcategorize for

word	possible clausal complements
know	THATCOMP
vow	THATCOMP, TOCOMP
eat	–
want	TOCOMP
resume	INGCOMP

Figure 5: Five verbs and the possible clausal complements they take based on the Dow Jones corpus

occurrences of the verb. [2]

We chose corpus analysis for transitivity determination, because information about transitivity in dictionaries is of little use. In a random sample of 50 verbs from the American Heritage Dictionary, 30 have both transitive and intransitive readings, e.g. *murder*. In everyday use (as well as in the MUC domain texts), the word is used almost exclusively as a transitive. Based on the MUC texts, the word "murder" gets a 0.85 transitivity rating.

We put the transitivity rating into each lexical item and use it to weight the likelihood of different parses. For example, whenever the grammar rule that builds a VP from a V and an NP fires, the weight of that VP is directly proportional to the verb's transitivity rating.

Automatically Acquiring Semantic Knowledge

Automatically acquiring semantic knowledge from text is more difficult than acquiring syntactic knowledge and generally requires a larger amount of text. However, we have been able to determine particular semantic features using syntactic clues.

Mass/Count nouns: SOLOMON's lexical entries distinguish between mass and count nouns. This is a semantic distinction that has syntactic and morphological correlates: count nouns pluralize much more frequently; they also take indefinite articles in the singular. We use these sorts of surface facts about the syntactic behavior of count and mass nouns to distinguish them and so provide the necessary information for our lexicons.

As with other phenomena, using statistical techniques offers a good way to collect data for individual nouns. The columns in Figure 6 represent various crucial statistics for predicting the mass or count nature of a given noun. The first column shows the raw frequency for singular occurrences, the second the raw frequency for plurals. The third represents the ratio of the two.

A high ratio of singular to plural uses is generally a strong indicator of the mass nature of the noun. However, there is at least one factor that introduces some noise into the data: when nouns occur in nonfinal position in a stacked noun phrase (i.e., a noun phrase consisting of several consecutive nouns), they are not inflected for number, as in *computer manual*. Such occurrences of singular nouns (or, strictly speaking, nouns that are neutralized for number) do not appear to even out, since some nouns seem to have a strong

[2]Our definition of transitivity is a heuristic. For example, we do not distinguish between those occurrences of verbs which are truly intransitive and those where the object is deleted due to a general syntactic process. Because instances of the latter are rare, they seem to have little effect on the transitivity rating.

	Plural Rating With Number	Plural Rating With Of
Random Sample of Words:		
LEFT	–	–
HOLDER	-2.9	3.1
INTRUSION	–	–
STRESS	–	–
MEDIATOR	–	–
BUREAU	–	–
SHOE	–	-0.2
SYNDROME	–	–
DEDICATION	–	–
LAUNDRY	–	–
BLITZ	–	–
ENOUGH	–	–
RESPONSIBILITY	–	–
DEFINITION	–	2.6
MEMO	–	–
RESURGENCE	–	–
CROP	–	–
BASEBALL	–	–
HELP	–	–
METHANE	–	–
Known Partitives:		
MILE	4.0	1.0
INCH	4.1	–
POUND	4.3	3.0
KILOGRAM	4.6	3.6
PART	0.0	3.3
PIECE	2.2	3.9
AMOUNT	4.8	–

Definitions of Column Headings

Plural Rating With Number: Mutual Association Ratio of any number with word

Plural Rating With Of: Mutual Association Ratio of word with "of"

Figure 7: Partitive Characteristics of a Random Selection of Nouns and Some Known Partitives from the Dow Jones Corpus.

	Singular Occs	Plural Occs	Occ Ratio	Occs With Directly Following Nouns	Occ Ratio Without Directly Following Nouns	Rating With *A* or *An*	Rating With Number
Mass nouns:							
furniture	60	0	–	26	–	1.8	–
software	358	0	–	149	–	0.7	–
water	256	33	7.758	108	4.845	-0.5	–
sodium	2	0	–	2	–	–	–
fat	21	8	2.625	3	2.250	1.0	–
food	527	61	8.639	316	3.459	0.4	–
harm	19	0	–	0	–	–	–
money	1700	0	–	412	–	-1.0	–
Count nouns:							
dog	29	23	1.261	10	0.826	3.6	–
telephone	321	18	17.833	270	2.833	2.3	–
emotion	14	10	1.4	0	1.400	–	–
computer	897	428	2.1	683	0.500	2.0	-1.5
book	278	165	1.685	69	1.267	2.4	0.0
animal	50	77	0.649	35	0.195	–	-1.6
Ambiguous nouns:							
cake	8	0	–	4	–	2.4	–
experience	212	16	13.25	8	12.75	0.3	–
feeling	50	37	1.351	1	1.324	2.4	–
sound	47	10	4.7	12	3.500	0.9	–
lamb	2	0	–	0	–	4.4	–
iron	24	2	12.0	16	4.000	5.0	–

Definitions of Column Headings

Singular Occs: Times word appeared tagged as a singular noun
Plural Occs: Times word appeared tagged as a plural noun
Occ Ratio: (Single Occs / Plural Occs)
Occs With Directly Following Nouns: Times singular form was followed directly by a noun
Occ Ratio Without Directly Following Nouns: (Single Occs – Occs With Nouns) / Plural Occs)
Rating With A or An: Mutual Association Ratio of *a* or *an* with the singular form of the word
Rating With Number: Mutual Association Ratio of any number with the plural form of the word

Figure 6: Mass/Count Characteristics of a Random Selection of Nouns from the Dow Jones Corpus

tendency to occur in that position. To illustrate, we give in the fourth column of Figure 6 the raw figures for frequency of occurrence of nouns standing immediately before another noun. When compared with the first column, some of the nouns show interesting skewing: *computer* shows a ratio of 683:897, *telephone* 270:321, *animal* 35:50. That some nouns should behave this way is resistant to explanation, but probably is related to the general nature of such nouns (there are lots of things pertaining to computers, hence noun phrases modified by *computer* are quite common).

We have simply compensated for this phenomenon in the count/mass statistics by only counting as singular nouns those occurrences of such in final position in a noun phrase. The fifth column in Figure 6 represents this "adjusted" ratio of singular to plural occurrences, leaving out the singular nouns occurring in nonfinal position in a stacked noun phrase. This column shows intuitively good results for mass versus count nouns (high versus low ratios).

The bottom of Figure 6 shows a class of nouns that are sometimes count and sometimes mass. We have termed them "ambiguous." As might be expected, their adjusted singular/plural ratios (fifth column) exhibit a mixed set of ratios (from the very high 12.75 for *experience* to the low 1.324 for *feeling*.

The second-to-last column of Figure 6 shows the Mutual Association Ratios for the selected set of nouns with a preceding indefinite article. The numbers match our intuitions: the mass nouns show a lower rate of occurrence with the article than the count nouns. Again, the ambiguous nouns exhibit mixed results.

The last column of Figure 6 shows that numbers occasionally occur preceding count nouns (e.g., *three computers*), but that they never occur preceding mass nouns or what we have defined as ambiguous nouns.

Partitives: Partitives are measure amounts: "three quarts," "two inches," etc. We have used the fact that partitive nouns are much more often preceded by numbers or followed by "of" phrases than other nouns so as to automatically detect them in context. Determining partitivity allows SOLOMON to better understand sentences such as "John has 3 buckets of sand." In the case of most noun phrases, the type of the corresponding semantic entity depends on the type of the head noun, as in "leader of the pack." This is not generally true of partitives. In "3 buckets of sand," the object of the "of" phrase determines the semantic type of the whole noun phrase, and is analyzed that way by SOLOMON.

Figure 7 contains some sample data for partitives derived from the Dow Jones corpus.

Conclusion

The work described above has established the value of statistically acquired knowledge in improving the performance of a knowledge-based system like SOLOMON. Statistical techniques allow the efficient and accurate scoping of the special characteristics of a new domain, where manual techniques are slow and subject to bias. They also have contributed directly to improving SOLOMON's performance.

Acknowledgments

We would like to thank Andrew Kehler for his collaboration on the research reported in this paper, as well as for his comments on an earlier draft.

References

Brent, Michael 1991. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of 29th Annual Meeting of the ACL.*

Church, Kenneth and Hanks, Patrick 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1).

Fano, R. 1961. *Transmission of Information: A Statistical Theory of Communications.* MIT Press, Cambridge, MA.

Hindle, Donald and Rooth, Mats 1991. Structural ambiguity and lexical relations. In *Proceedings of 29th Annual Meeting of the ACL.*

Kehler, Andrew; Blejer, Hatte R.; Flank, Sharon; and McKee., Douglas 1990. A three-tiered parsing approach for operational systems. In *Proceedings of the AI Systems in Government Conference.*

Shugar, Shelton; Kehler, Andrew; Flank, Sharon; Blejer, Hatte; and Maloney, John 1991. Language independence in Project MURASAKI. Presented at the Eighth Annual Intelligence Community AI Symposium.

Smadja, Frank and McKeown, Kathleen 1990. Automatically extracting and representing collocations for language generation. In *Proceedings of 28th Annual Meeting of the ACL.*

Tomita, Masaru 1986. *Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems.* Kluwer, Boston, MA.